INNOVATIONS VI
In Cancer Prevention and Research Conference

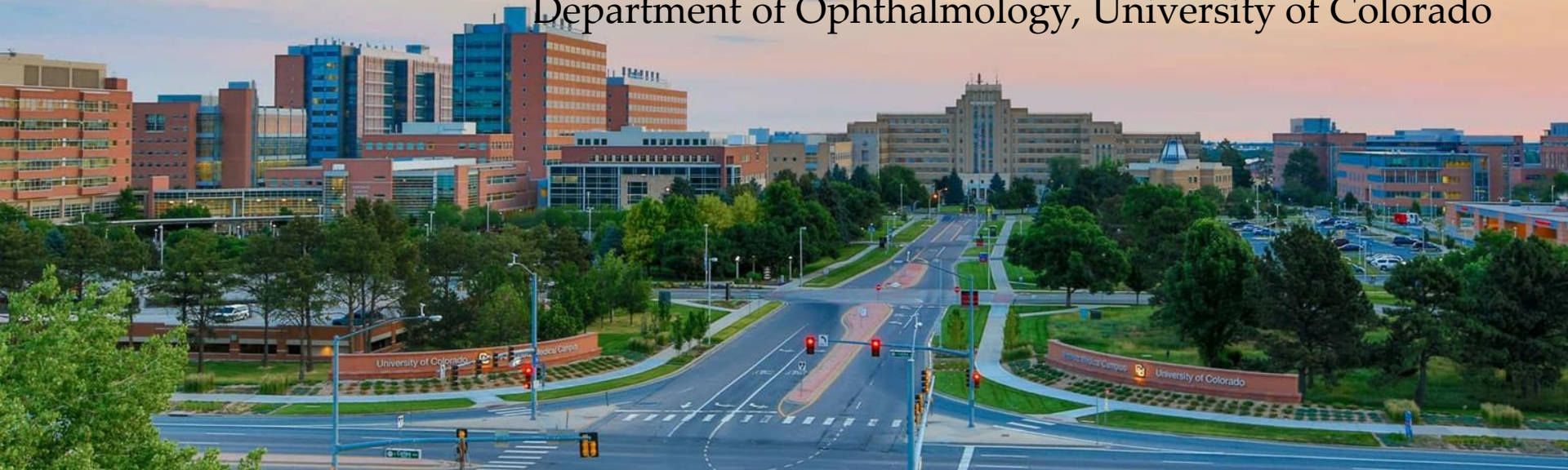
CANCER PREVENTION & RESEARCH
INSTITUTE OF TEXAS

Early Detection: Advances in Artificial Intelligence in Imaging

University of Colorado **Anschutz Medical Campus**

Developing Imaging AI Biomarkers in Oncology: Opportunities and Challenges

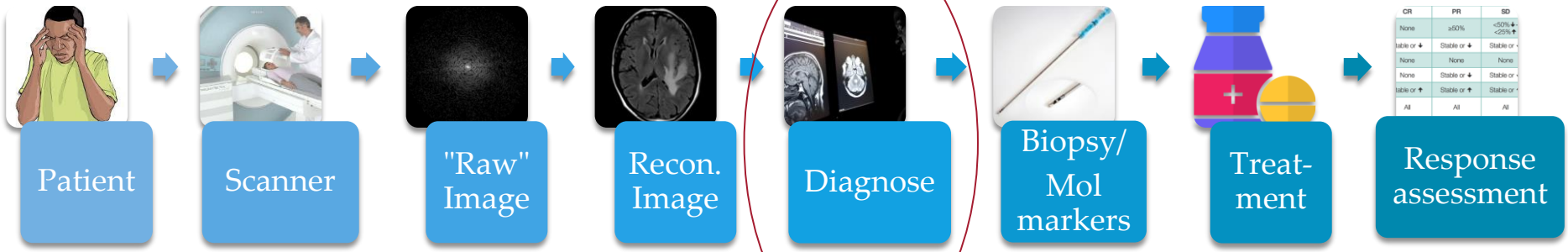
Jayashree Kalpathy-Cramer, PhD
Chief, Division of Artificial Medical Intelligence
Department of Ophthalmology, University of Colorado



Disclosures

- Deputy editor, Radiology-AI
- Research support from Genentech
- i-ROP DL FDA breakthrough status licensed to Boston AI lab
- Grant support from NIH, NSF, EU
- Consultant, Siloam Vision Inc.

AI/ML is being used widely throughout the entire patient journey



AI can be applied throughout the workflow

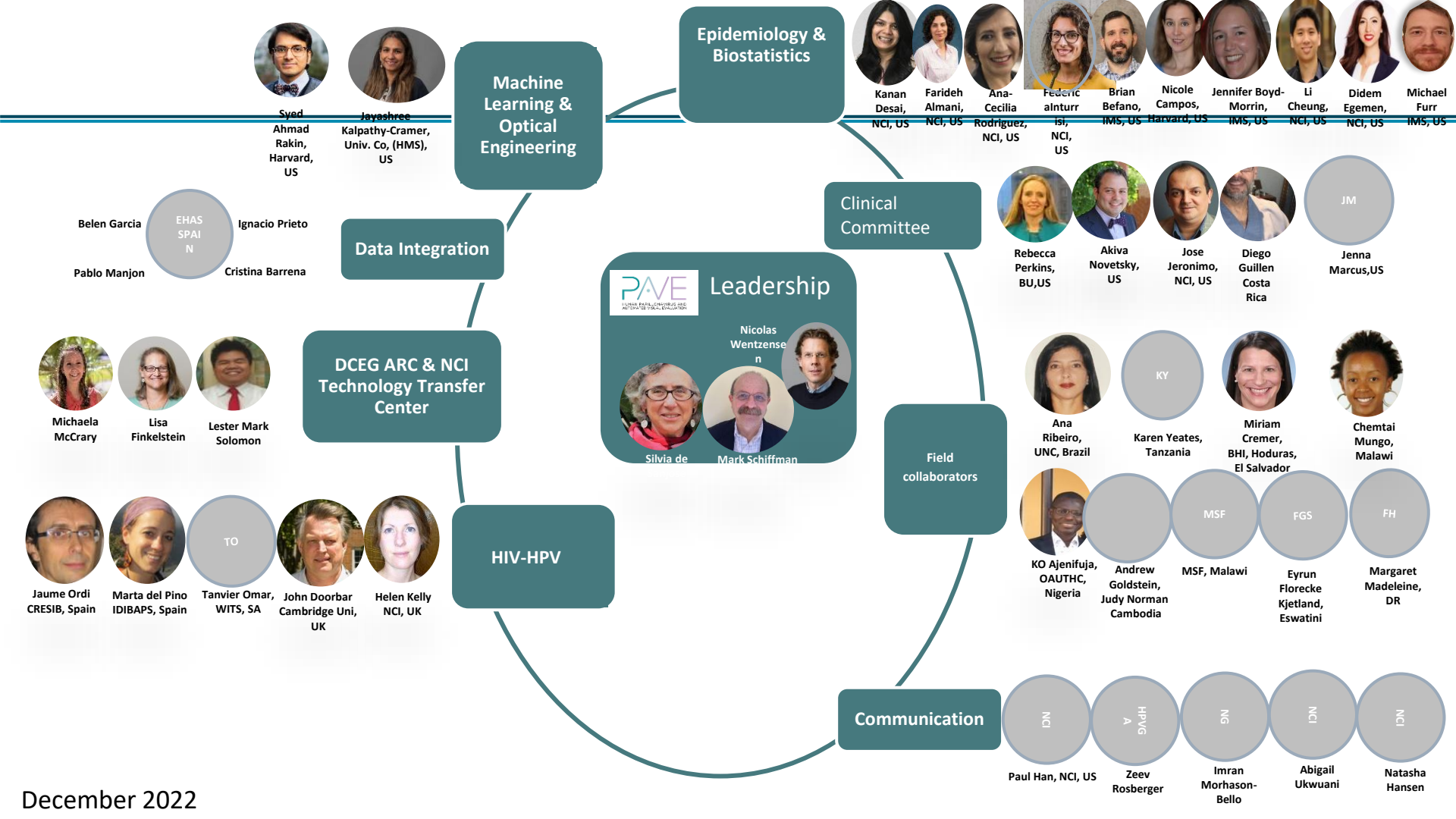
Cervical Cancer Screening (collaboration with NCI)

Cervical cancer is a leading cause of cancer morbidity and mortality worldwide

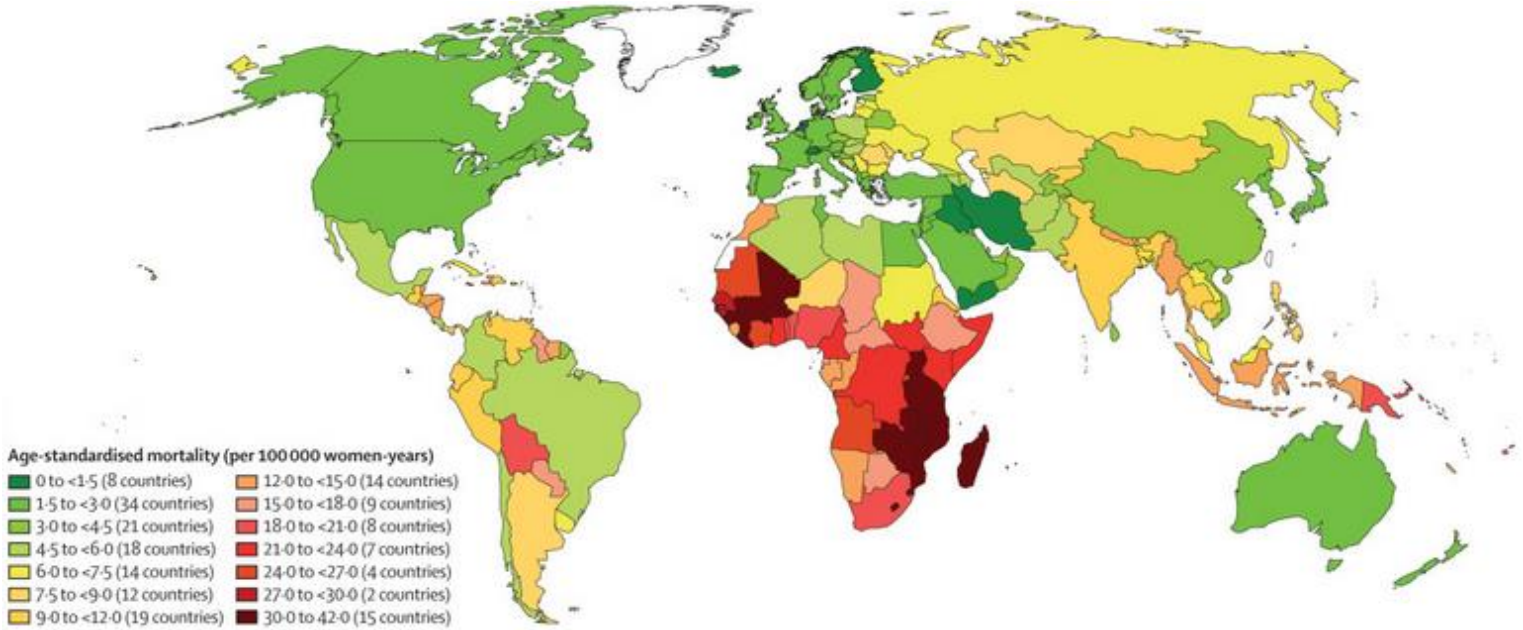
Persistent infections with high-risk human papilloma virus (HPV) strains remain the strongest risk factor for subsequent neoplastic growth

Screening of the cervix by visual inspection after application (VIA) of acetic acid

ML algorithm for analysis of images of the cervix, in conjunction with HPV testing can be used in global screening programs



December 2022



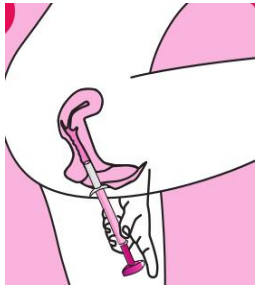
Sources: Arbyn, Marc, et al. The Lancet Global Health 8.2 (2020): e191-e203.

MOONSHOT INITATIVE

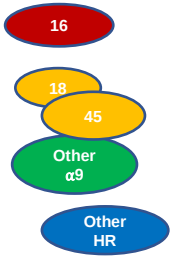
PAVE

HUMAN PAPILLOMAVIRUS AND AUTOMATED VISUAL EVALUATION

Candidate single step screening + primary triage



Self-sampling



HPV extended genotyping

Candidate secondary triage

AI-based Automated visual evaluation with a mobile image capture device

Desai K. et al. Infect Agent Cancer 2020

Combined with treatment

Battery-operated point-of-care ablation devices

Mobile LLETZ devices

Slide: PAVE team

PAVE risk stratification

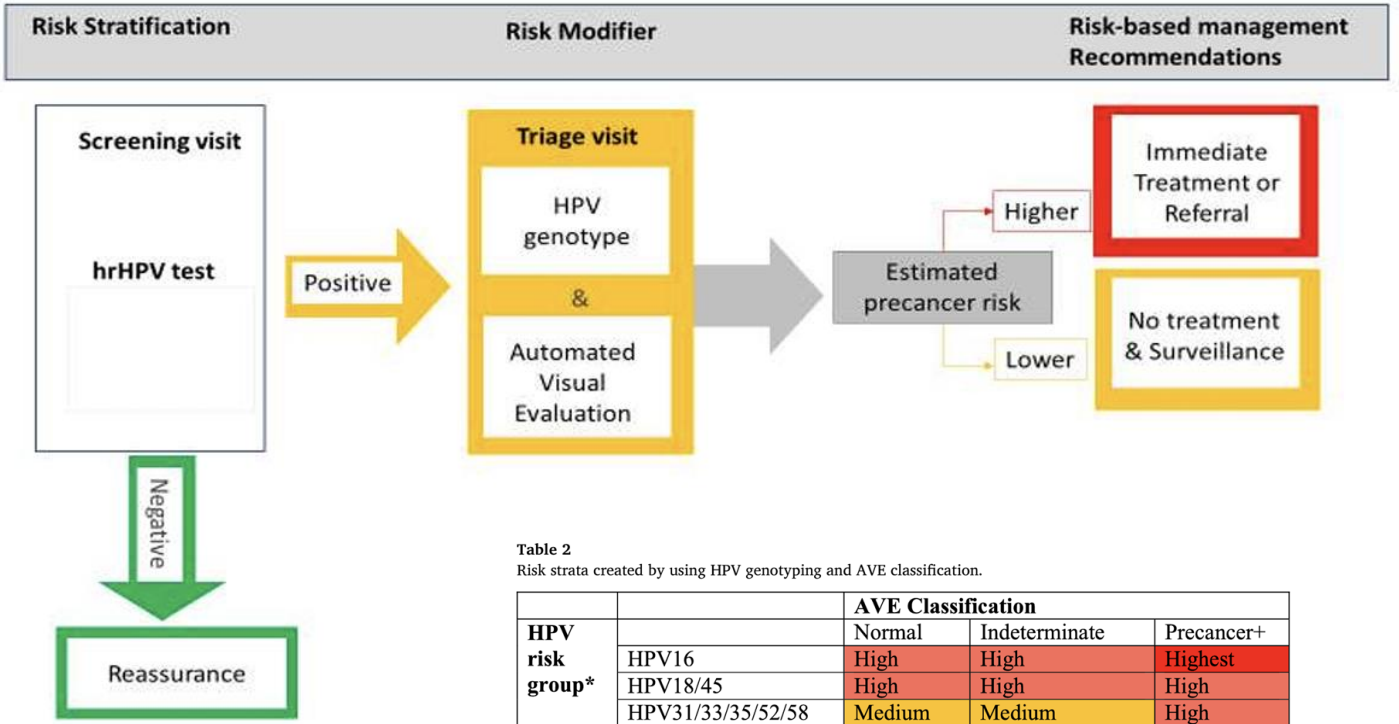


Table 2
Risk strata created by using HPV genotyping and AVE classification.

		AVE Classification		
HPV risk group*		Normal	Indeterminate	Precancer+
	HPV16	High	High	Highest
	HPV18/45	High	High	High
	HPV31/33/35/52/58	Medium	Medium	High
	HPV39/51/56/59/68	Low	Medium	High

*In case of multiple infections, the result will be hierarchical, as HPV16 else HPV18/45 else HPV31/33/35/52/58 else HPV39/51/56/59/68. For analyses with limited numbers, the two middle categories (HPV18/45 group and HPV 31/33/35/52/58 group) can be combined, leading to a three-part scale (HPV16, intermediate, low). The expectation is that the ordinality of the scale will remain constant across settings but the absolute risk of precancer+ may vary by population characteristics.

Perkins et

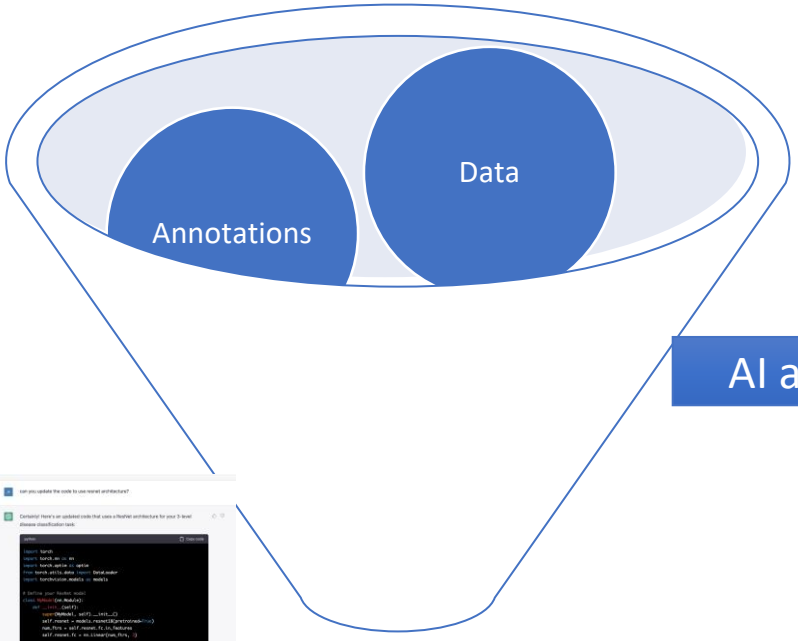
REPRODUCIBLE AND CLINICALLY TRANSLATABLE DEEP NEURAL NETWORKS FOR CERVICAL SCREENING

Syed Rakin Ahmed[1,2,3,4,†], Brian Befano[5,6,†], Andreanne Lemay[1,7], Didem Egemen[8], Ana Cecilia Rodriguez[8], Sandeep Angara[9], Kanan Desai[8], Jose Jeronimo[8], Sameer Antani[9], Nicole Campos[10], Federica Inturrisi[8], Rebecca Perkins[11], Aimee Kreimer[8], Nicolas Wentzensen[8], Rolando Herrero[12], Marta del Pino[13], Wim Quint[14], Silvia de Sanjose[8,15], Mark Schiffman[8], Jayashree Kalpathy-Cramer[1]

Assessing generalizability of an AI-based visual test for cervical cancer screening

Syed Rakin Ahmed[1,2,3,4,†], Didem Egemen[5,†], Brian Befano[6,7], Ana Cecilia Rodriguez[5], Jose Jeronimo[5], Kanan Desai[5], Carolina Teran[8], Karla Alfaro[9], Joel Fokom-Domgue[10,11,12], Kittipat Charoenkwan[13], Chemtai Mungo[14], Rebecca Luckett[15], Rakiya Saidu[16], Taina Raiol[17,18], Ana Ribeiro[17,18], Julia C. Gage[19], Silvia de Sanjose[5,20], Jayashree Kalpathy-Cramer[1,21,*], Mark Schiffman[5,*]

It is becoming *really easy* to create an AI algorithm today!



AI algorithm

AI Algorithm

Title

Performance of an AI algorithm for the diagnosis of retinopathy of prematurity

Abstract

Retinopathy of prematurity (ROP) is a potentially blinding eye disorder that affects premature infants. Early diagnosis and treatment of ROP is critical for preserving vision and preventing blindness. In recent years, artificial intelligence (AI) algorithms have shown promise for the automated diagnosis of ROP. In this study, we evaluated the performance of an AI algorithm for the diagnosis of ROP using a dataset of fundus images from premature infants.

We trained a convolutional neural network (CNN) on a dataset of 5,000 fundus images from premature infants with and without ROP. We evaluated the performance of the CNN on a separate dataset of 1,000 fundus images from premature infants, including 500 images with ROP and 500 images without ROP. We measured the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy of the CNN for the diagnosis of ROP.

Our results showed that the CNN achieved a sensitivity of 95.2%, a specificity of 93.8%, a PPV of 92.1%, an NPV of 96.2%, and an accuracy of 94.5% for the diagnosis of ROP. The area under the receiver operating characteristic curve (AUC-ROC) was 0.96, indicating excellent diagnostic accuracy.

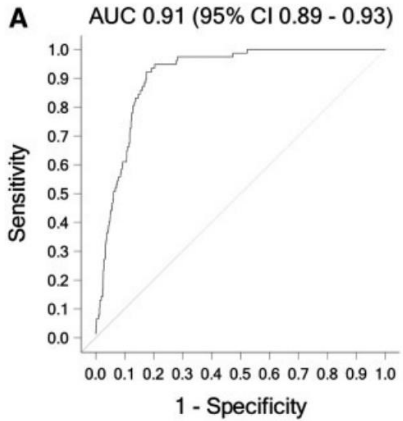
Our study demonstrates that an AI algorithm based on a CNN can achieve high diagnostic accuracy for the diagnosis of ROP. The use of AI algorithms for the automated diagnosis of ROP has the potential to improve the efficiency and accuracy of ROP screening programs, particularly in resource-limited settings where access to ophthalmologists and specialized equipment may be limited.

Cervical cancer screening using deep learning for AVE

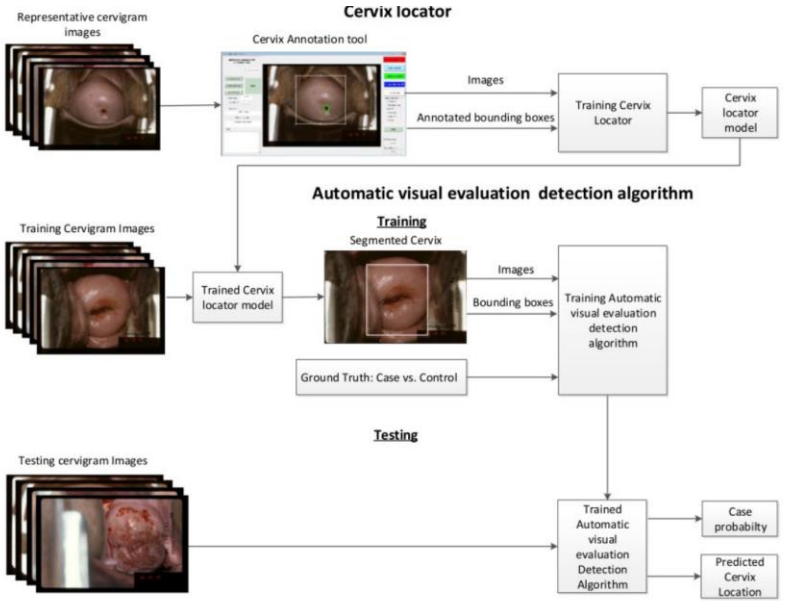
ARTICLE

An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening

Liming Hu, David Bell, Sameer Antani, Zhiyun Xue, Kai Yu, Matthew P. Horning, Noni Gachuhi, Benjamin Wilson, Mayoore S. Jaiswal, Brian Befano, L. Rodney Long, Rolando Herrero, Mark H. Einstein, Robert D. Burk, Maria Demarco, Julia C. Gage, Ana Cecilia Rodriguez, Nicolas Wentzensen, Mark Schiffman



"Automated visual evaluation of enrollment cervigrams identified cumulative precancer/cancer cases with … accuracy 0.91"



Follow up publication in cervical cancer highlighting some of the challenges

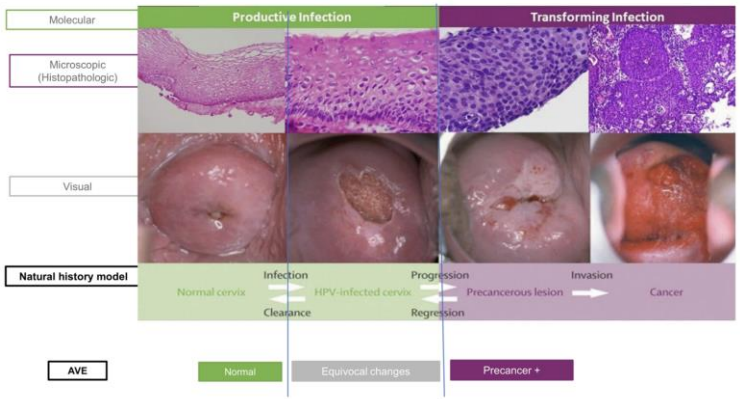
Received: 7 July 2021 | Revised: 24 September 2021 | Accepted: 15 October 2021
DOI: 10.1002/ijc.33879

SPECIAL REPORT

IJC INTERNATIONAL JOURNAL of CANCER

The development of "automated visual evaluation" for cervical cancer screening: The promise and challenges in adapting deep-learning for clinical testing

Kanan T. Desai[1] | Brian Befano[2,3] | Zhiyun Xue[4] | Helen Kelly[1] | Nicole G. Campos[5] | Didem Egemen[1] | Julia C. Gage[1] | Ana-Cecilia Rodriguez[1] | Vikrant Sahasrabuddhe[6] | David Levitz[1] | Paul Pearlman[7] | Jose Jeronimo[1] | Sameer Antani[4] | Mark Schiffman[1] | Silvia de Sanjosé[1,8]



2 | STEP-WISE CONSIDERATIONS FOR AI-BASED AVE

2.4 | Validation of the output of the algorithm

2.4.1 | Reproducibility of AVE

2.4.2 | Internal validity of AVE

2.4.3 | External validity (generalizability) of AVE and avoiding overfitting

2.4.4 | Device portability of AVE

2.4.6 | Risk prediction: "calibration" of AVE

2.4.7 | Predicting immediate vs future risk

J Natl Cancer Inst. 2023 Sep 27;djad202. doi: 10.1093/jnci/djad202. Online ahead of print.

AI-based image analysis in clinical testing: lessons from cervical cancer screening

Didem Egemen [1], Rebecca B Perkins [2], Li C Cheung [1], Brian Befano [3 4], Ana Cecilia Rodriguez [1], Kanan Desai [1], Andreanne Lemay [5], Syed Rakin Ahmed [5 6 7 8], Sameer Antani [9], Jose Jeronimo [1], Nicolas Wentzensen [1], Jayashree Kalpathy-Cramer [5], Silvia De Sanjose [1 10], Mark Schiffman [1]

Affiliations + expand

PMID: 37758250 DOI: 10.1093/jnci/djad202 Sign in

Lessons learned

1) Specify rigorously what the algorithm is designed to identify and what the test is intended to measure, e.g., screening, diagnostic, or prognostic.
2) **Design the AI algorithm to minimize the most clinically important errors.**
3) **Evaluate AI algorithms like any other test, using clinical epidemiologic criteria.**
4) Link the AI algorithm results to clinical risk estimation.
5) Generate risk-based guidelines for clinical use that match local resources and priorities.

Challenges in “real life” AI deployment

- **Generalizability– models are brittle and do not generalize across scanners, populations, disease presentation**
- **Model predictions may not be repeatable!**
- **Gray zone” -many diseases lie on a spectrum, ratings are binary/ordinal**
- **Calibration- commonly used approaches for binary models can lead to poorly calibrated models**
  - Silent failures – models may fail without indication (”confidently wrong”)
- **Overfitting – reported model performance can be over-optimistic**
- Explainability
- Models can be biased (in hard to detect ways)

Problem 1: “Brittleness” of machine learning models

Deep learning models do not generalize well

Only 6% of published AI studies have external validation (Kim et al., KJR, 2019)

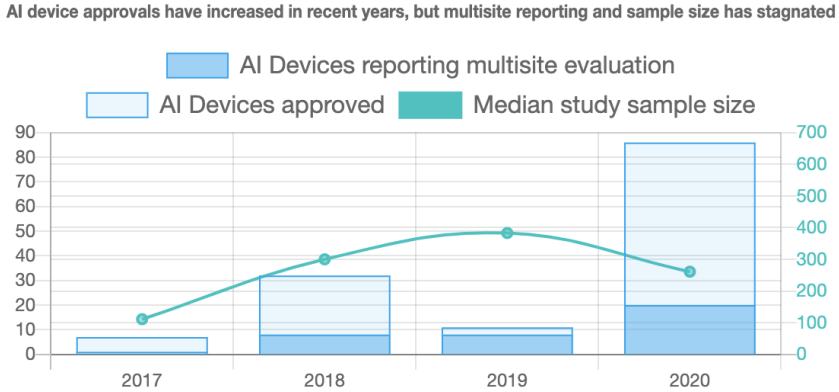
Data heterogeneity can lead to poor model performance on external datasets.

Few FDA approved AI devices have been evaluated externally

How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals

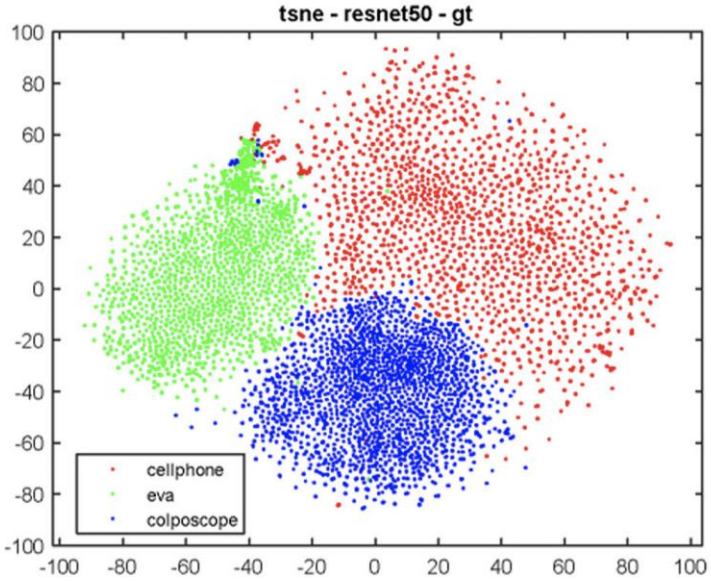
A comprehensive overview of medical AI devices approved by the US Food and Drug Administration sheds new light on limitations of the evaluation process that can mask vulnerabilities of devices when they are deployed on patients.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho and James Zou



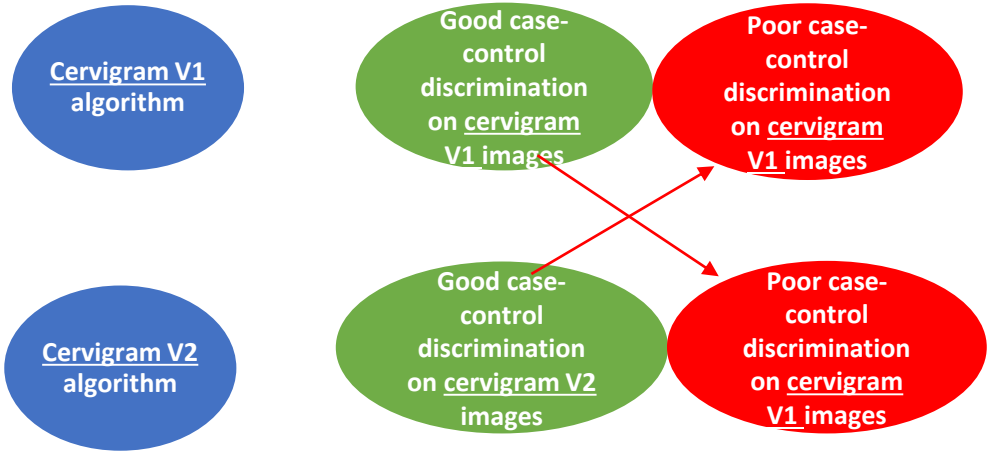
https://ericwu09.github.io/medical-ai-evaluation/

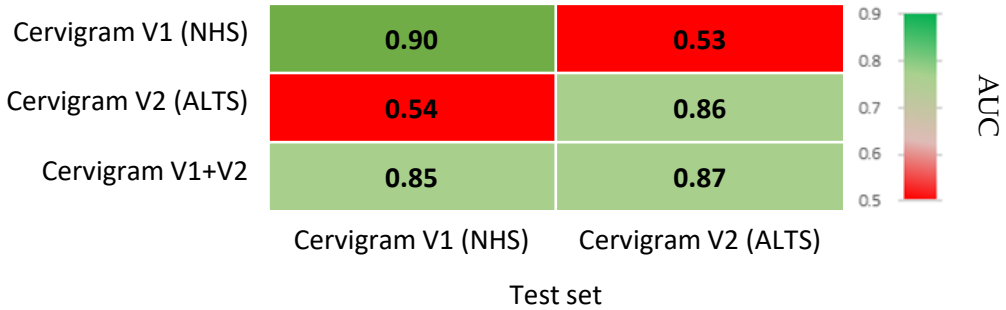
Images acquired on different devices can be quite different



Xue et al, 2021 doi:10.1109/ cbms52027.2021.00085

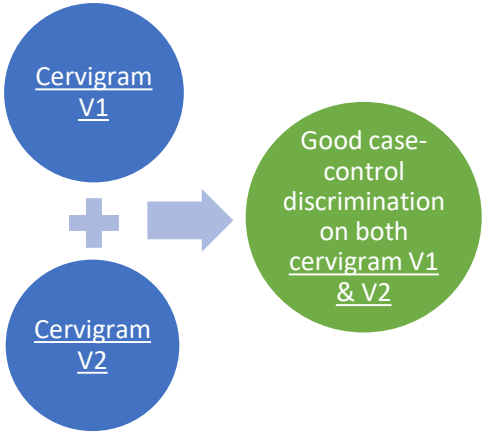
“Portability challenges” in cervical cancer





ning set

Combined model



Slide: KananDesai/Mark Schiffman

Solution 1: WIP!

Increase data diversity

Multi-institutional databases

Out of distribution detection

Novel AI methods to improve generalization

Federated learning

Self-supervised learning

Problem 1: Evaluation plan

Curate multi-institutional or multi-scanner datasets

Consider ways in which “out-of-distribution” input may occur

- Different scanners
- Poor quality
- Wrong anatomy/modality/view
- Different demographics (e.g. pediatric)

Evaluate performance on unseen datasets

Continuous monitoring

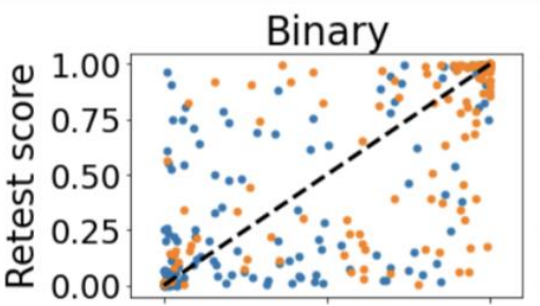
Problem 2: DL Model predictions are not repeatable!

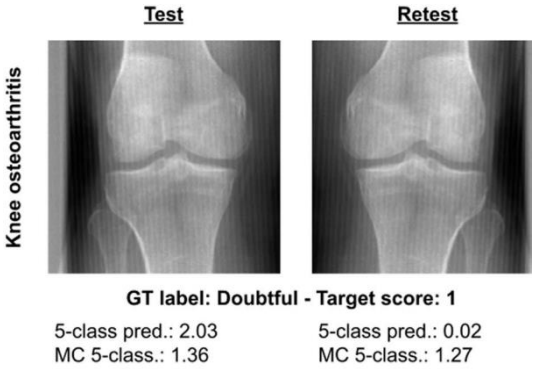
A replicate set of images yield different results (lack of repeatability)

(a) Model prediction: 0.01 (Normal)

(b) Model prediction: 0.98 (Pre-cancer)

Fig. 1: Illustration of repeatability issues from deep learning models on different images of a cervix with precancerous lesions from the same patient taken the same day. A binary model without dropout layers generated the following outputs. (a) the binary model predicts a normal cervix (severity score: 0.01), (b) the binary model predicts pre-cancer (severity score: 0.98).





Little published literature on model repeatability/reproducibility

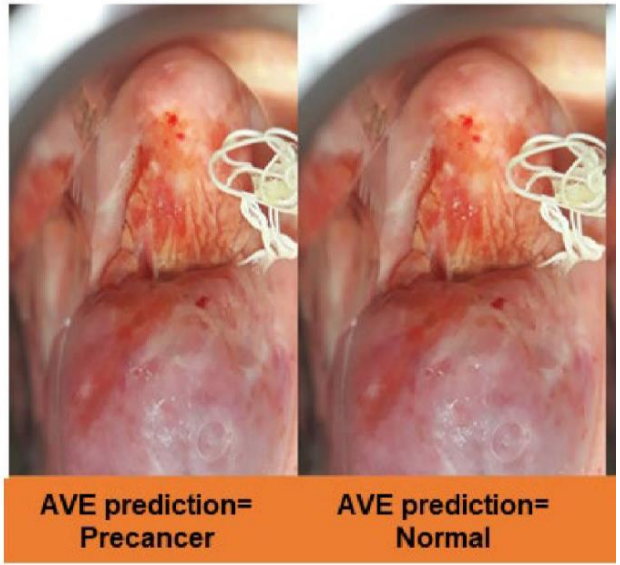
Many models are not repeatable when tested!.

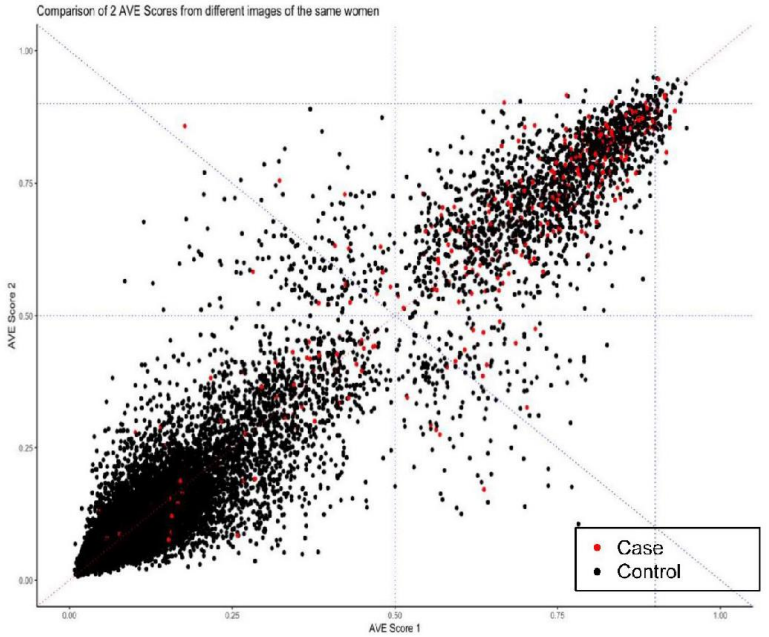
Desai K et al. IJC 2021;

Lemay et

Problem 1: Test-retest repeatability can be an issue

Challenge: *A replicate set of images from a woman during same examination with same device, yielded different results* ***(lack of repeatability)***





This issue was

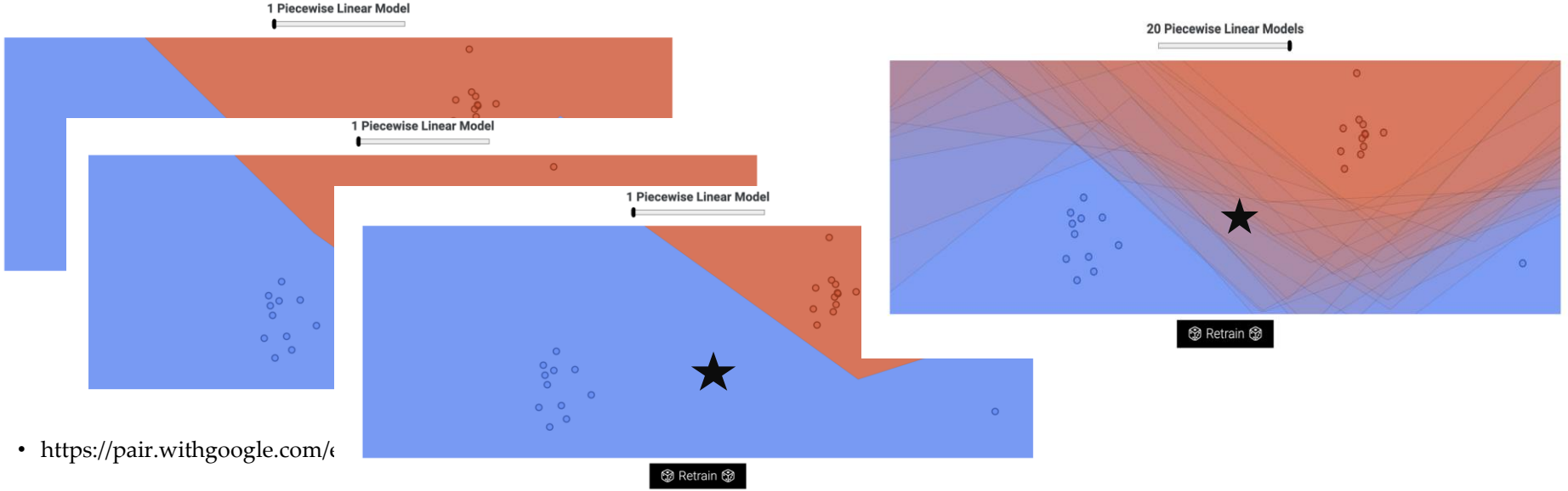
Desai K et al. IJC 2021;

PAIR EXPLORABLES

From Confidently Incorrect Models to Humble Ensembles

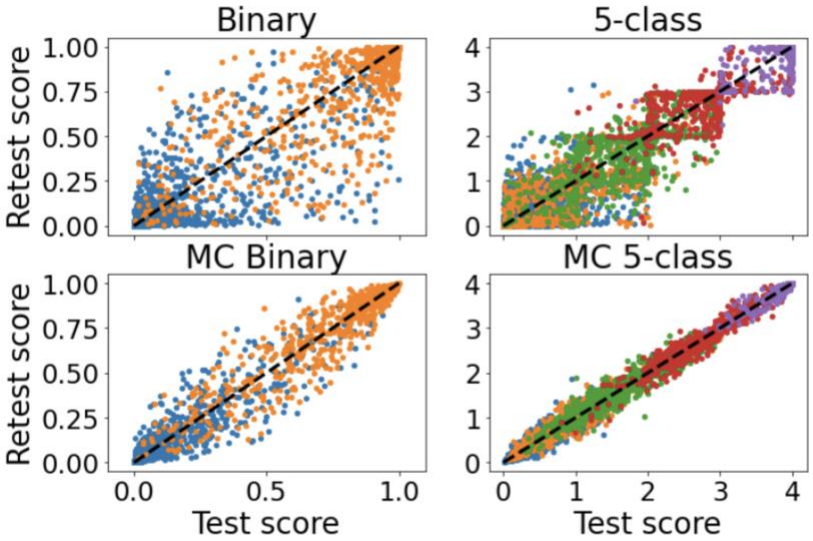
Combining Models Reduces Overconfidence

By averaging the output of multiple models, a technique known as **ensembling**, we can create a model

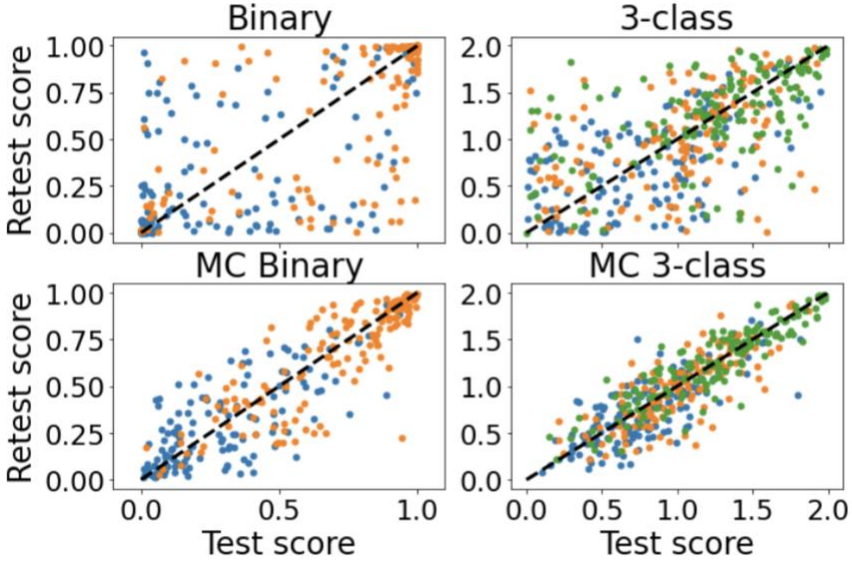


- https://pair.withgoogle.com/e

Solution 1: Monte Carlo approaches may improve repeatability



Knee osteoarthritis (xray)



Cervical cancer (photos)

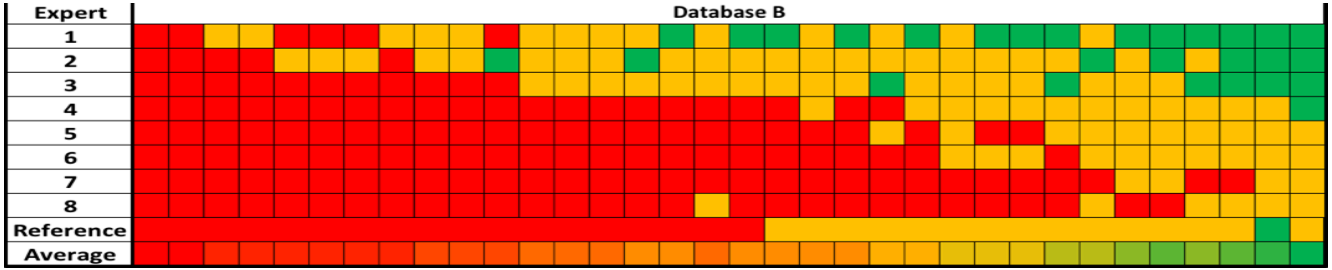
Lemay et al. NPJ

Problem 2: Evaluation plan

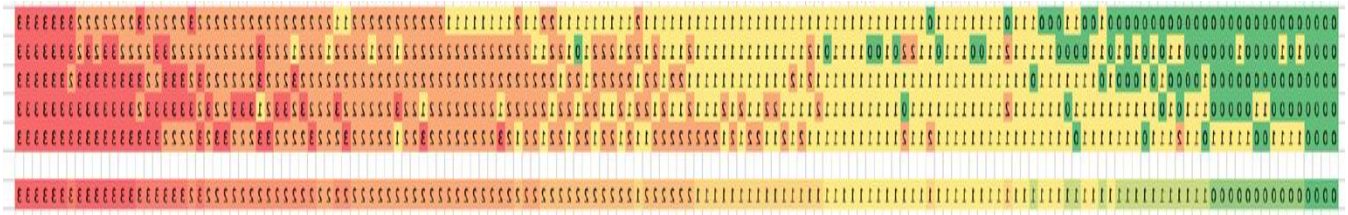
Curate datasets to evaluate repeatability/reproducibility

- If ethical, acquire test-retest datasets of the same patients
- If not, generate datasets with slight variations (e.g. flip image, rotate image slightly)

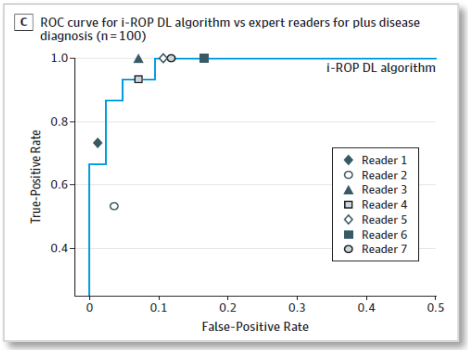
Problem 3: Real World is a Often a Continuous Spectrum



Campbell et al, Ophthalmology 2016;123:2338-44.



Li et al, Academic Radiology, 2021



Problem 3 Diseases lie on a spectrum

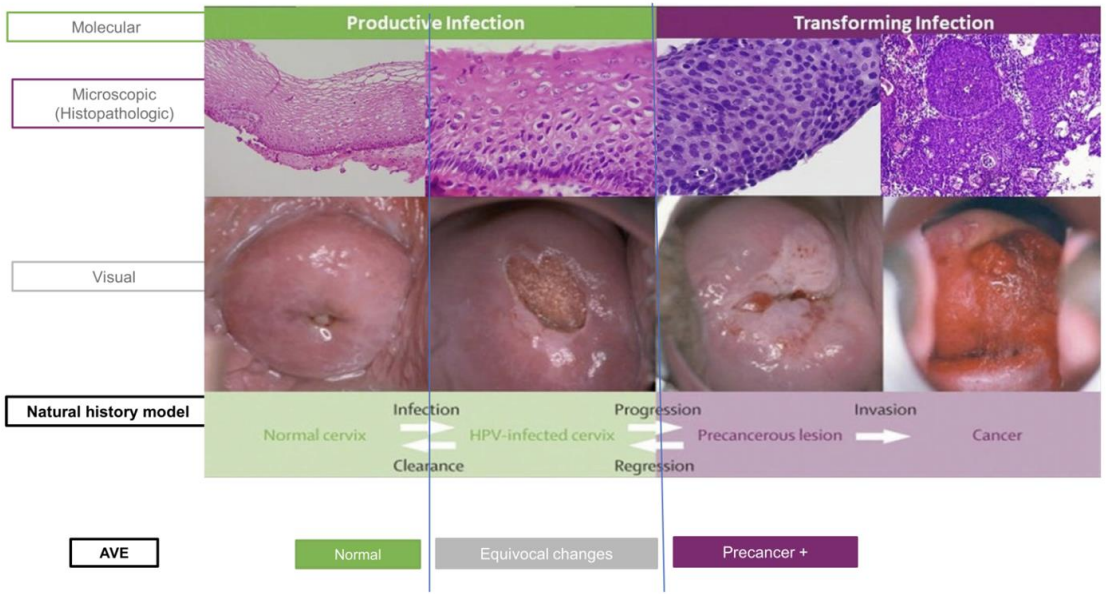


FIGURE 3 The AVE classification categories are expected to be consonant with the four biological distinct stages in the natural history and pathogenesis of cervical cancer. Reprinted with permission from Schiffman et al[21]; Histopathology image source: Desai et al[20] [Color figure can be viewed at wileyonlinelibrary.com]

Not recognizing equivocal changes can lead to extreme mis-classifications and grave errors

Slide:
Desai et

Problem 3: DL model may have extreme misclassifications/confidently wrong

Challenge: Distinguishing HPV related equivocal changes from precancer is challenging leading to **extreme misclassification** by binary AVE classifier

		Classification	
		Normal	Precancer+
Ground Truth	Normal	93.7%	6.3%
	Precancer+	33.6%	66.4%

Solution 3: Introduce a “gray zone”

<u>Solution:</u> Adding equivocal class in training a three-class ordinal classifier reduced serious misclassification

		Classification		
		Normal	Equivocal	Precancer+
Ground Truth	Normal	**83.7%**	**13.4%**	**2.9%**
	Equivocal	**35.8%**	**30.2%**	**34%**
	Precancer+	**8.6%**	**11.4%**	**80%**

Solution 3: Generate continuous output variables instead of binary values

npj | Digital Medicine

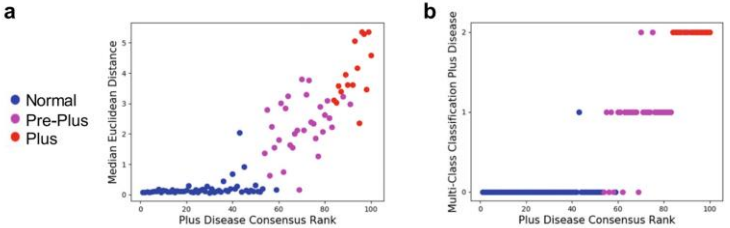
www.nature.com/npjdigitalmed

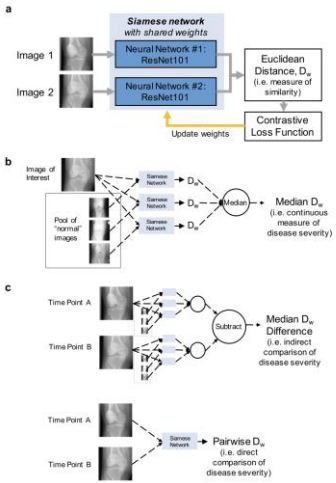
ARTICLE OPEN

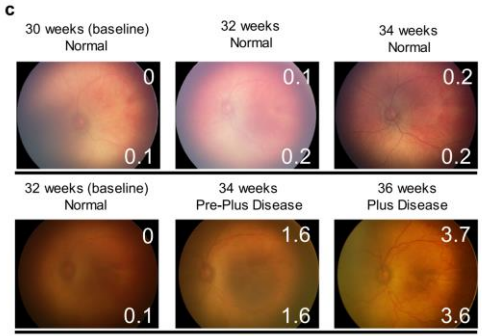
Check for updates

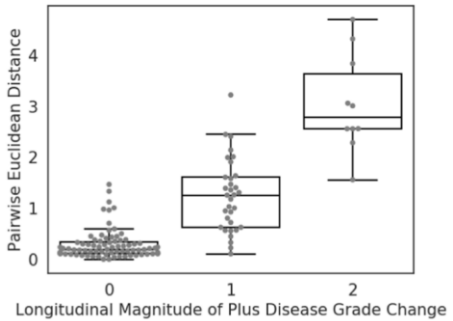
Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging

Matthew D. Li[1], Ken Chang[1], Ben Bearce[1], Connie Y. Chang[2], Ambrose J. Huang[2], J. Peter Campbell[3], James M. Brown[4], Praveer Singh[1], Katharina V. Hoebel[1], Deniz Erdoğmuş[5], Stratis Ioannidis[5], William E. Palmer[2], Michael F. Chiang[3,6] and Jayashree Kalpathy-Cramer[1,7]✉









Problem 3: Evaluation plan

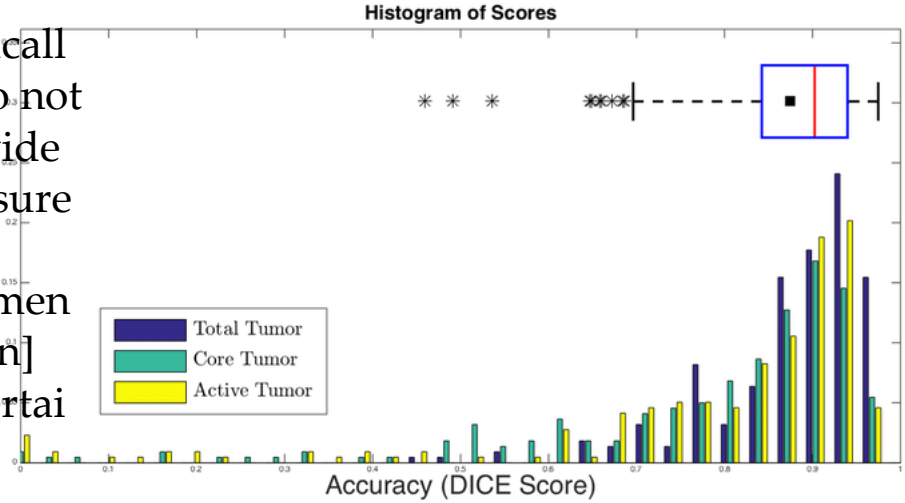
Generate datasets with multiple raters

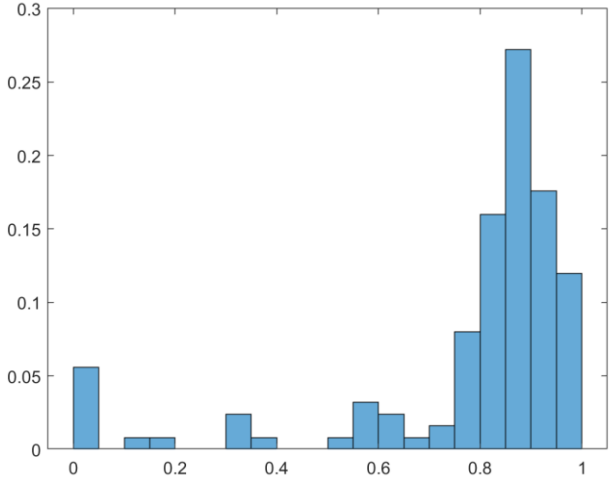
Generate datasets along disease spectrum

Evaluate (binary) models on nuanced cases

Problem 4: Models may fail silently

Deep learning approaches (typically) do not provide measures of [segmentation] uncertainty



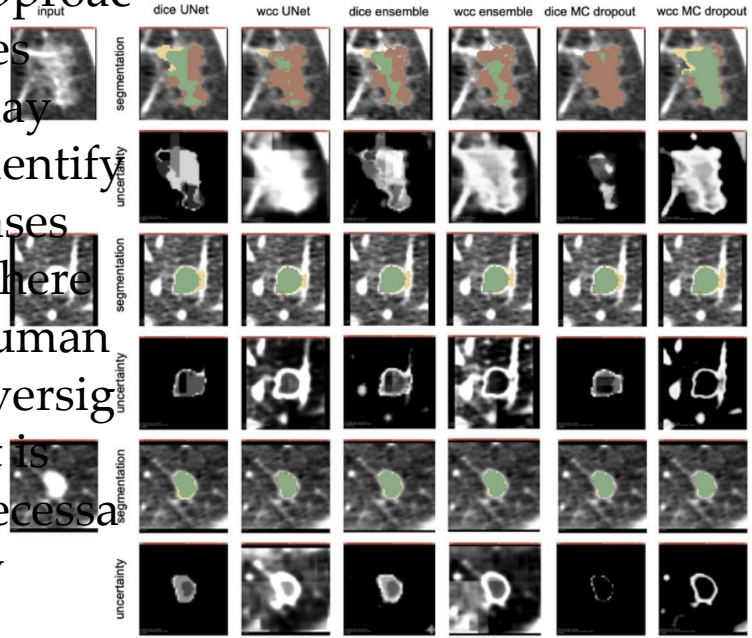


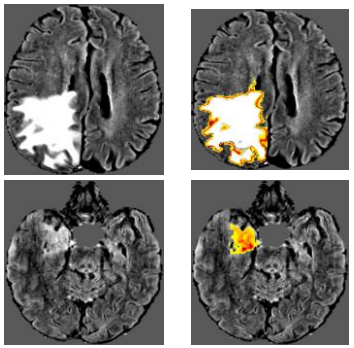
Example histogram of dice

3D Convolutional Neural Networks for Glioblastoma Segmentation
Darvin Yi, Mu Zhou, Zhao Chen, Olivier Gevaert

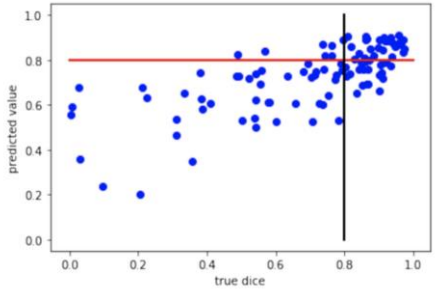
Solution 4: Bayesian DL approaches (Monte Carlo) may provide estimates of voxel and patient level uncertainty

Such approaches may identify cases where human oversight is necessary

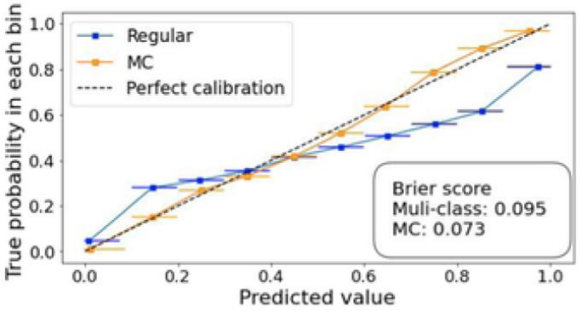




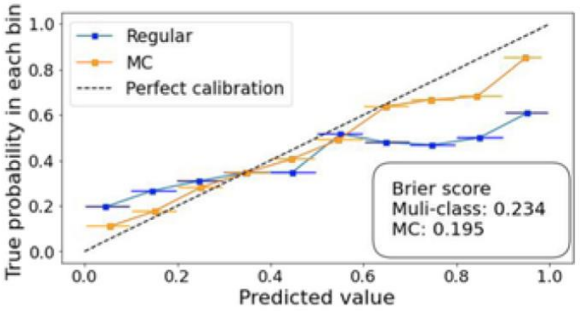
Linear regression model to predict the Dice score of a segmentation prediction (here test data set) from its uncertainty measures. The red line indicates the chosen cutoff for flagging of 0.8 and the red line the true decision boundary of 0.8.



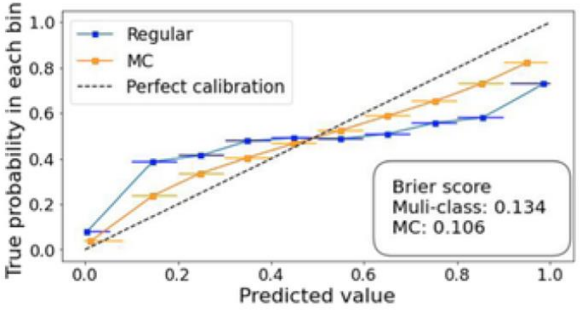
Solution 4: Methods such as MC may improve calibration



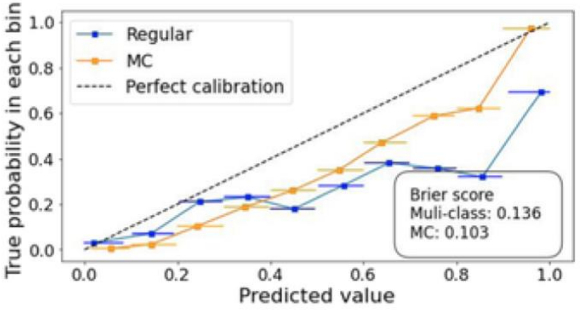
(a) Knee osteorarthritis 5-class models



(b) Cervical 3-class models



(c) Breast density 4-class models



(d) ROP binary models

Validation

Opinion | Open Access | Published: 24 February 2023

There is no such thing as a validated prediction model

Ben Van Calster, Ewout W. Steyerberg, Laure Wynants & Maarten van Smeden

BMC Medicine **21**, Article number: 70 (2023) | Cite this article

Reason 1: patient populations vary
Reason 2: measurements of predictors or outcomes vary
Reason 3: populations and measurements change over time

https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-023-02779-w

Problem 4: Evaluation plan

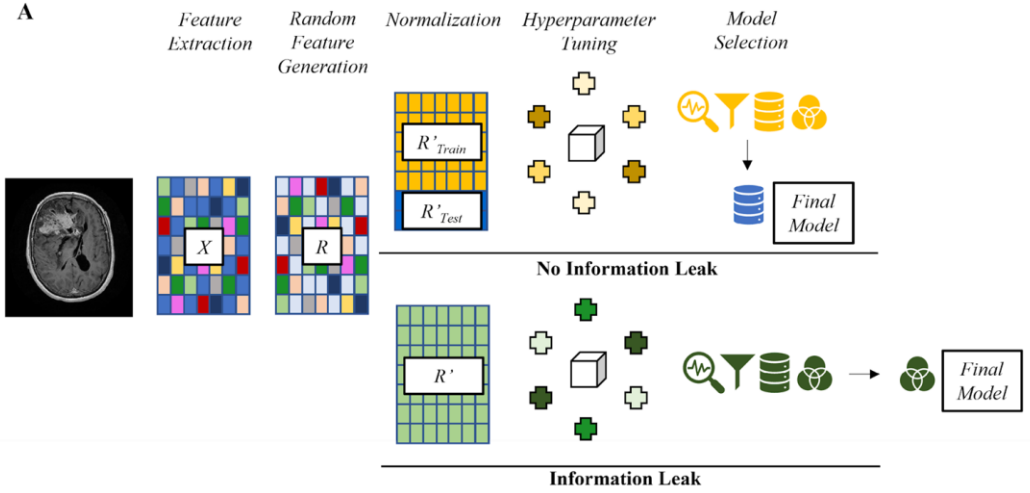
Evaluate model calibration

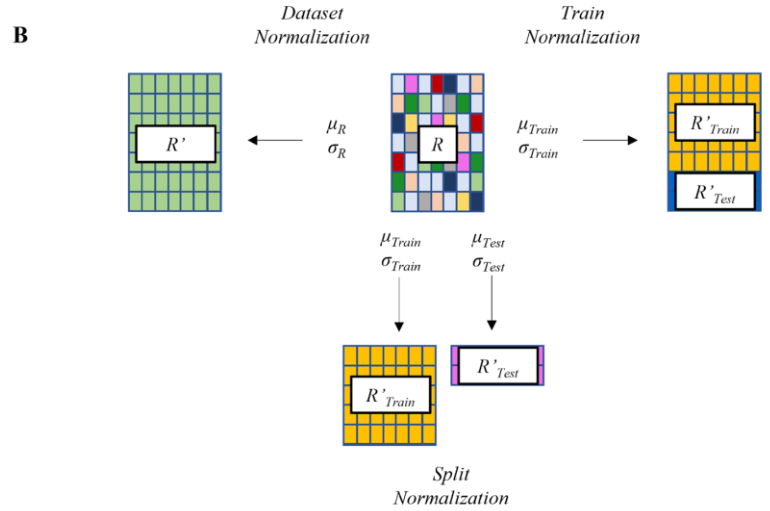
Tabulate silent failures and confidently wrong predictions

how often? Any defining characteristics?

Problem 5: Overfitting is a common problem in the literature

The literature is rife with over-optimistic reported performance, primarily due to a lack of statistical rigor.





Radiology

ORIGINAL RESEARCH · COMPUTER APPLICATIONS

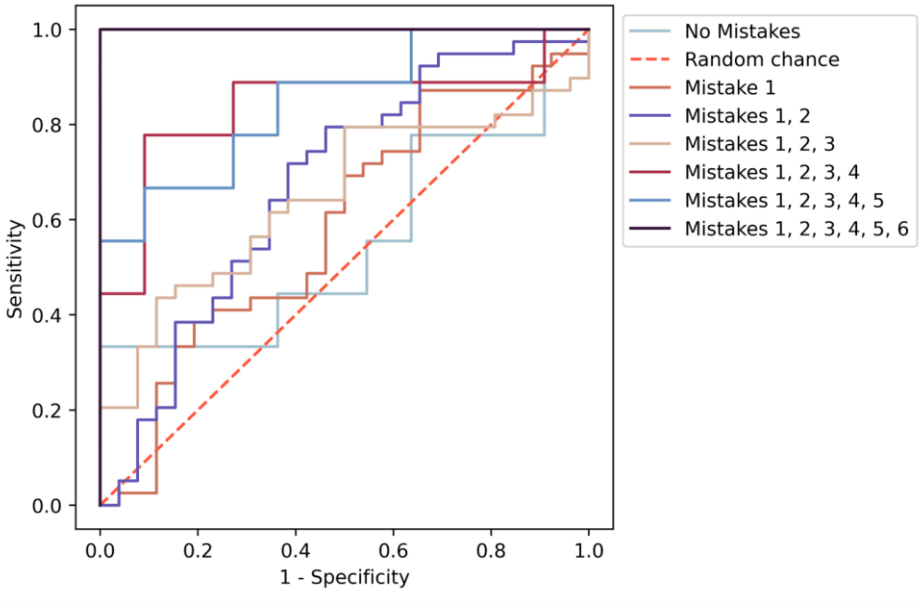
Inconsistent Partitioning and Unproductive Feature Associations Yield Idealized Radiomic Models

Mishka Gidwani, PhD • Ken Chang, MD, PhD • Jay Biren Patel, BS • Katharina Viktoria Hoebel, MD • Syed Rakin Ahmed, BA • Praveer Singh, PhD • Clifton David Fuller, MD, PhD • Jayashree Kalpathy-Cramer, PhD

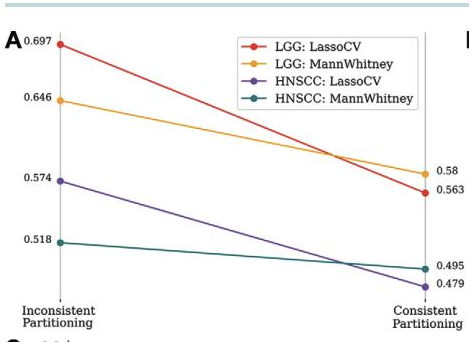
Gidwani et al,

Problem 5: Overfitting is a common problem in the literature

The literature is rife with over-optimistic reported performance, primarily due to a lack of statistical rigor.



Mistakes	Added mistake	ROC-AUC
None		0.556
1	Feature normalization in batch	0.585
12	Feature selection in batch	0.672
123	Model selection using test set	0.647
1234	No external validation set	0.838
12345	Hyperparameter selection in batch	0.848
123456	Report results on all data	1.0



Gidwani et al.

Solution 5: Best practices and statistical rigor throughout

Are the data representative of the population of interest?

Do the

Checklist before model deployment

- ✓ What is the reproducibility/ portability performance?
- ✓ What is repeatability (test-retest performance) of the model?
- ✓ Does the system have an “out of distribution” detector?
- ✓ How well is the model calibrated?
- ✓ How often does the model make grave errors? Is the model confidently wrong?
- ✓ Is the model explainable?
- ✓ Is the model biased? Fair?

- ✓ What is the continuous monitoring plan?

Thanks!!!

Yi-Fen Yen
Assistant Professor

Ina Ly
Research Fellow

Dania Daye
MD/PhD

Katharina Hoebel
Graduate Student

Chris Bridge
Instructor

Syed Rakin Ahmed
Graduate Student

Advaith Veturi
Senior Data Scientist

Dagoberto Pulido-Arias
Research Analyst

Jay Patel
Graduate Student

Benjamin Bearce
Senior Full Stack Web Developer

Praveer Singh
Post Doctorate

Albert Kim
Physician Investigator

Chris Clark
Data Scientist

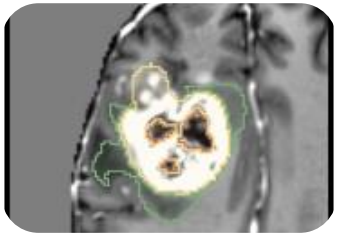
Mason Cleveland
Programmer Analyst

Satvik Tripathi
Research Student

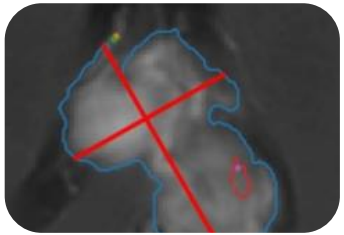
Randy Lu
Research Student

Thanks to funding from NIH, NSF, EU!

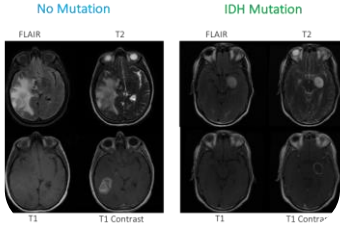
Deep Learning in Oncological Imaging



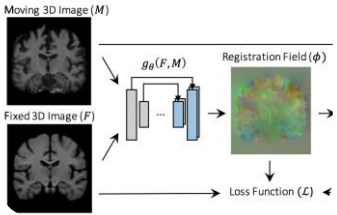
Segmentation



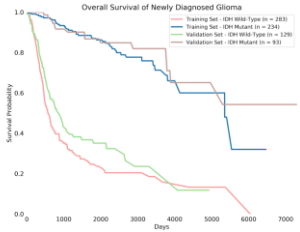
Response Assessment



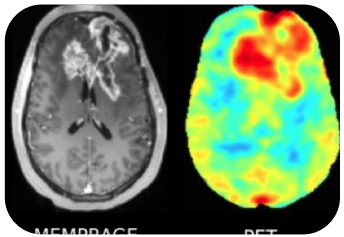
Radiogenomics



Registration



Survival Prediction

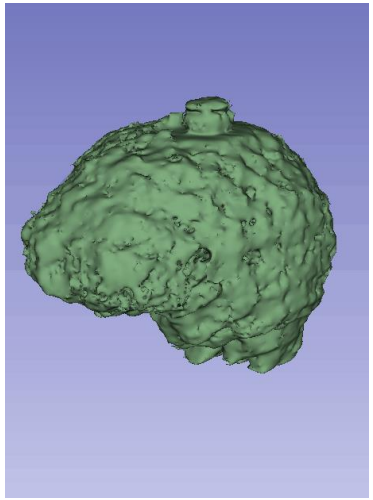


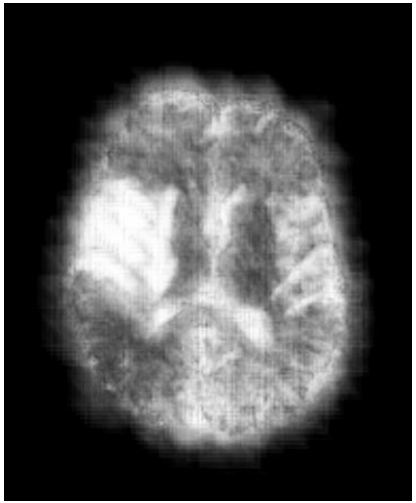
Drug delivery

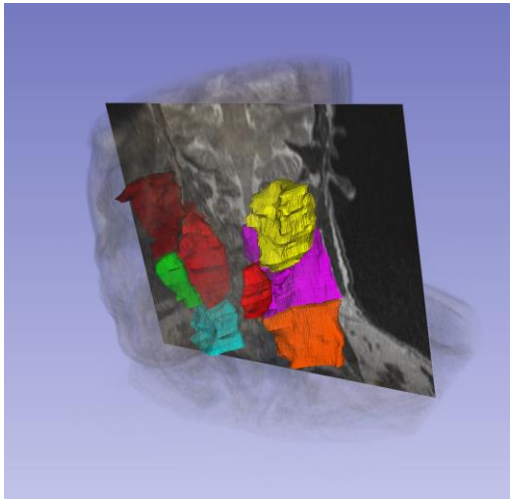
Segmentation (delineation of object boundary) is often used in oncology and radiation oncology

Quantifying tumor burden at a single time point and longitudinally

Contouring of tumors and organs at risk is key in radiation therapy planning

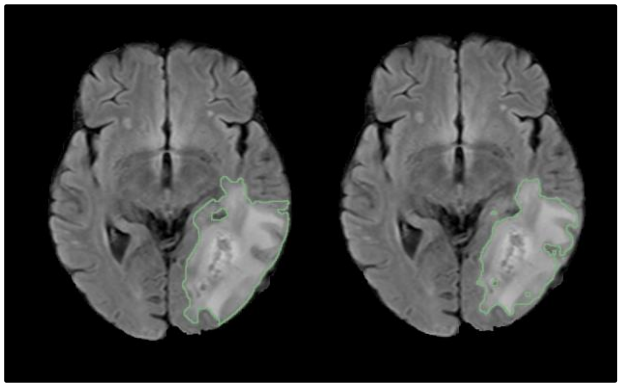




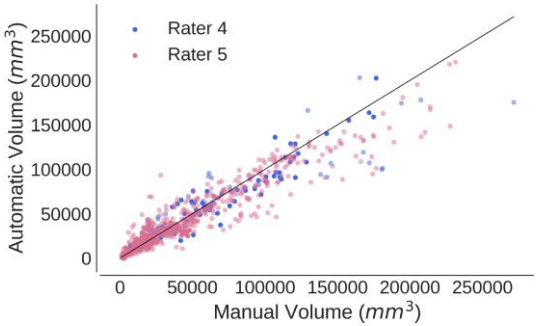


Tumor volume measurements agree with experts

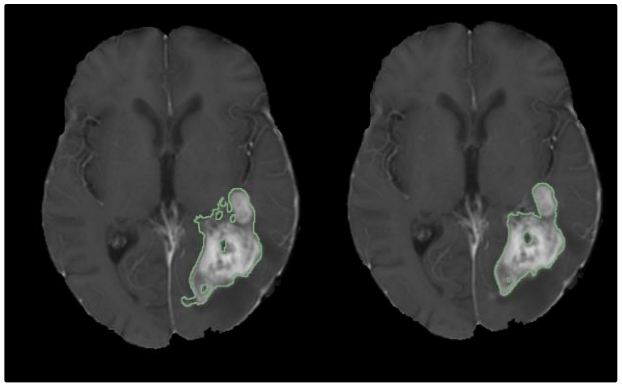
FLAIR Hyperintensity



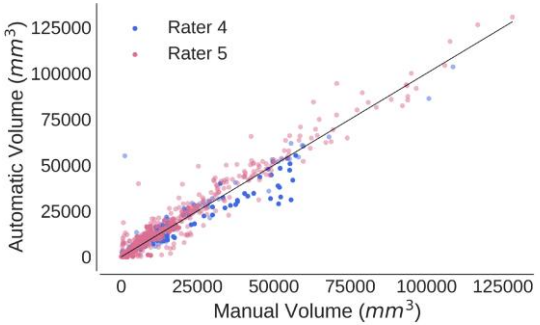
Manual Automatic



Enhancing Tumor

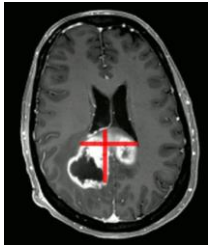


Manual Automatic



Chang, Beers*, Bai* et al., Neuro-Oncology (2019)*

Response Assessment: Is the patient responding to therapy?



Response Assessment in Neuro-Oncology (RANO)
Changes in bi-directional measurement of enhancing tumor

Wen et al., JCO (2010)
Reuter et al., J Neurooncol (2014)

Sounds easy enough!

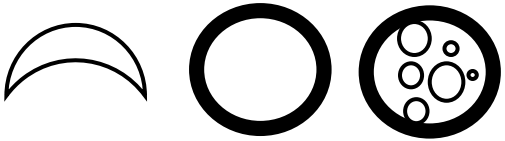
The step-by-step approach

1) Find the axial slice with largest tumor area
2) Find the largest measurable* diameter, excluding necrosis and blood
3) Find the largest measurable* perpendicular diameter
4) Multiply diameters
5) Repeat for up to 5 lesions and sum

→ Done visually

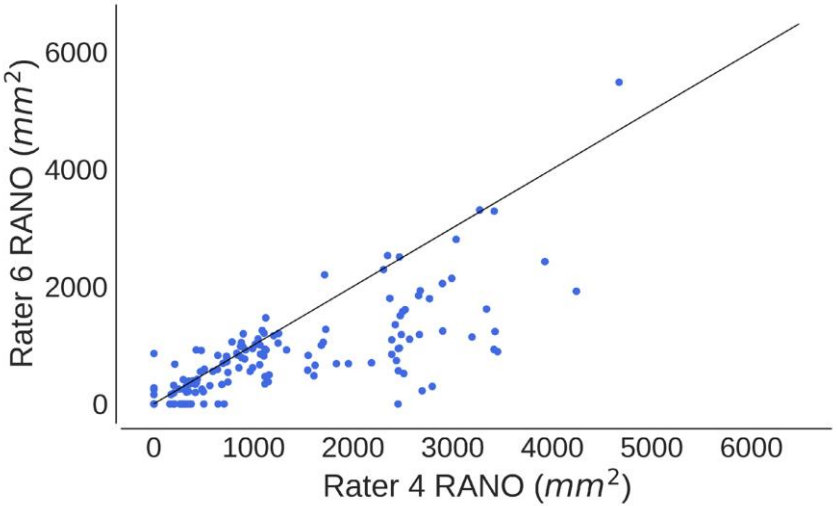
→ Depending on how you call necrosis/blood, tumor may or may not be measurable

→ Help! The tumor is an odd shape



→ Compounding of any variability in 1-4

Moderate agreement between clinicians



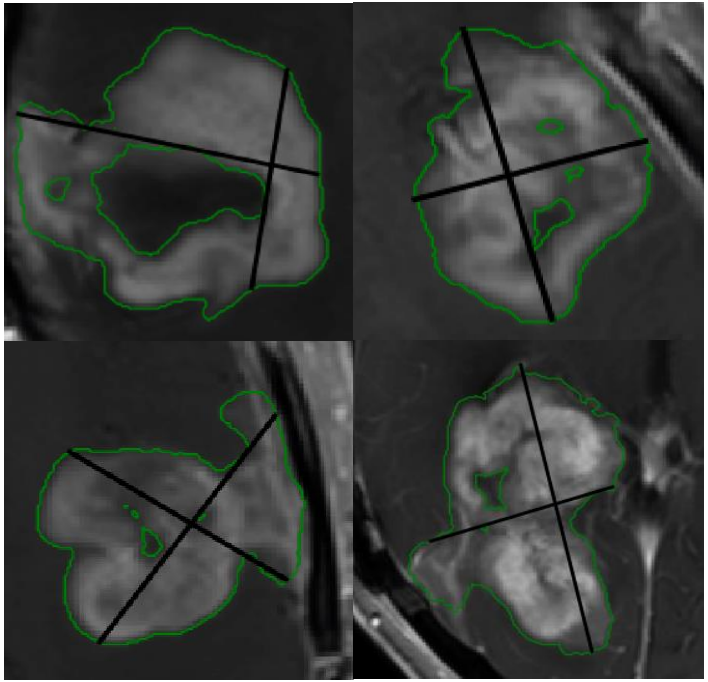
Why not just use volume?

Bi-directional measurements are easier than volume measurements (less time)!

Slide: Ken Chang

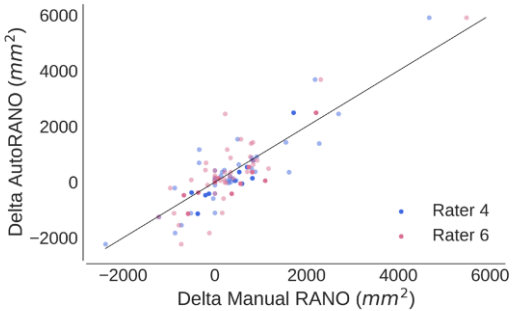
AutoRANO

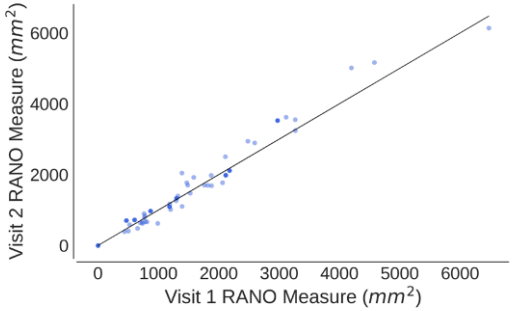
- Use the axial slice with the largest area (actually!)
- Because we use automatic segmentations, can consistently exclude blood and necrosis
- Diameters can exit segmentation for up to 10% of its length (to account for small holes)



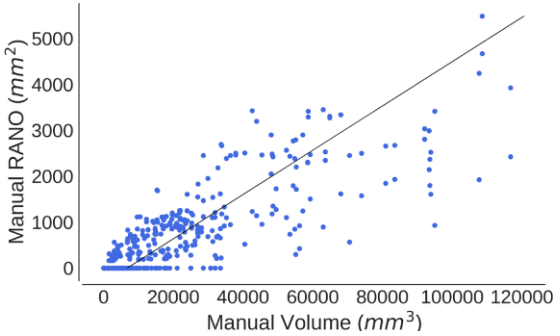
Slide: Ken Chang

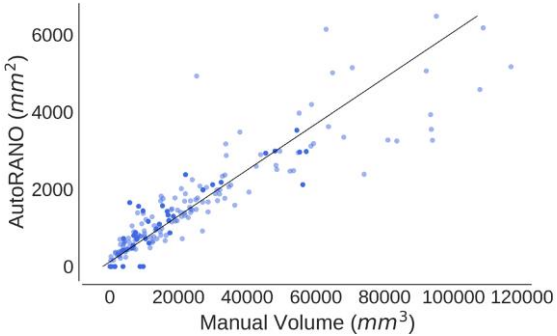
Performance of AutoRANO





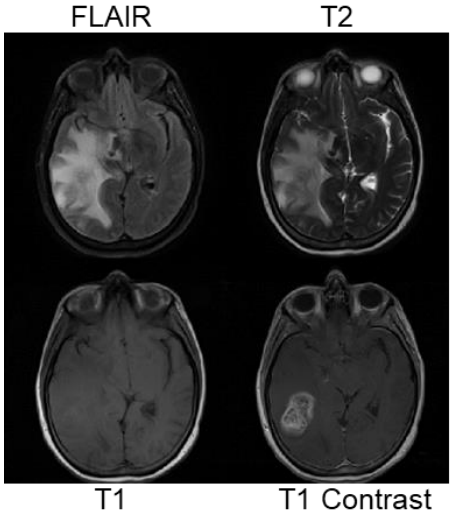
AutoRANO is more reflective of tumor volume than manual RANO



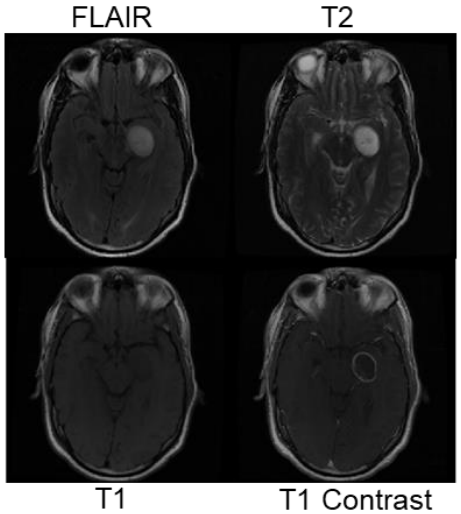


A “virtual” biopsy

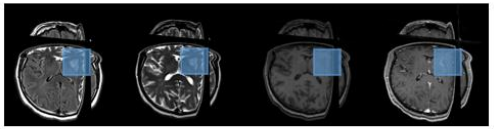
No Mutation

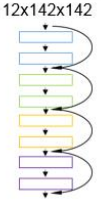


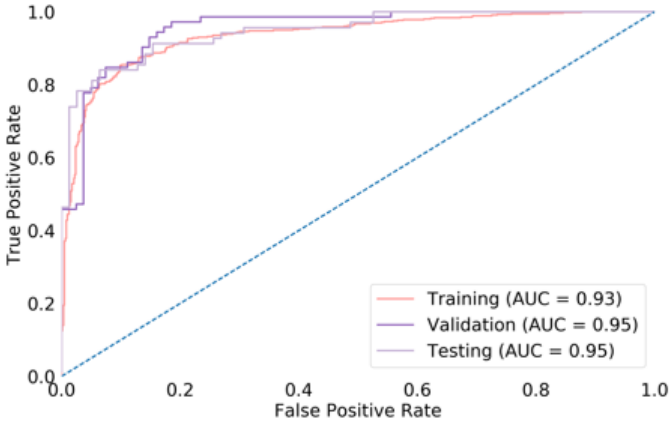
IDH Mutation



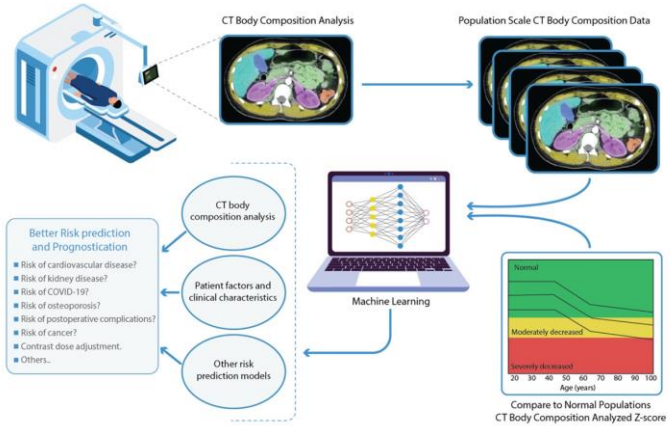
Less aggressive growth
Sharp margins
Homogenous signal intensity
Less contrast enhancement







Opportunistic Screening



Review

Role of Machine Learning-Based CT Body Composition in Risk Prediction and Prognostication: Current State and Future Directions

Tarig Elhakim [1,2,*], Kelly Trinh [3], Arian Mansur [4], Christopher Bridge [2,4] and Dania Daye [2,4,*]

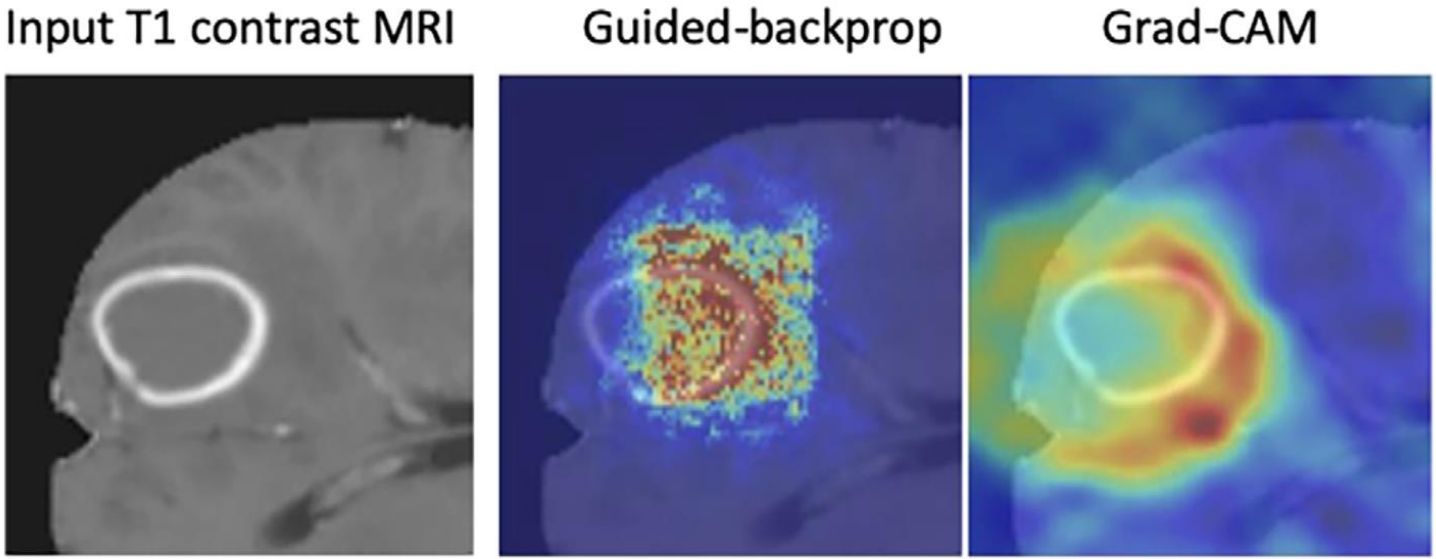
nature medicine

Article

https://doi.org/10.1038/s41591-023-02232-8

Body composition and lung cancer-associated cachexia in TRACERx

Post-hoc methods (saliency maps)



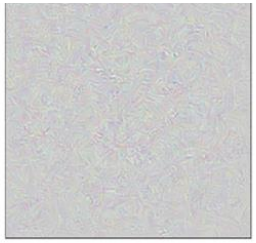
Reyes et al, Rad-AI 2020

Problem: Shortcut learning

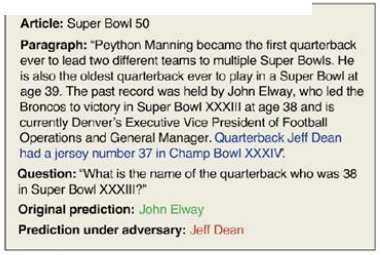
Shortcut learning in deep neural networks

Robert Geirhos[1,2,4] ✉, Jörn-Henrik Jacobsen[3,4], Claudio Michaelis[1,2,4], Richard Zemel[3,5], Wieland Brendel[1,5], Matthias Bethge[1,5] and Felix A. Wichmann[1,5]

Shane 2018



Zech 2018

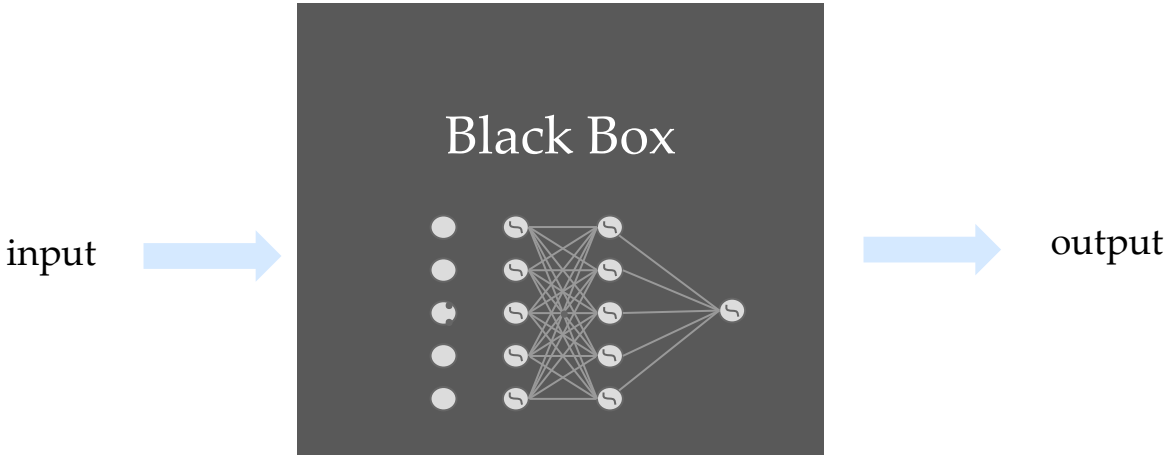


Jia 2017

Task for DNN	Caption image	Recognize object	Recognize pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognize primary object	Uses features unrecognizable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Fig. 1 | Examples of shortcut learning. Deep neural networks often solve problems by taking shortcuts instead of learning the intended solution, leading to a lack of generalization and unintuitive failures. This pattern can be observed in many real-world applications. Figure adapted with permission from ref. [14], AI Weirdness (left); ref. [17], PLOS (third from left).

Problem 6: Deep learning models can be black-boxes



Problem 6: Current explainability methods have limitations

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

The Mythos of Model Interpretability

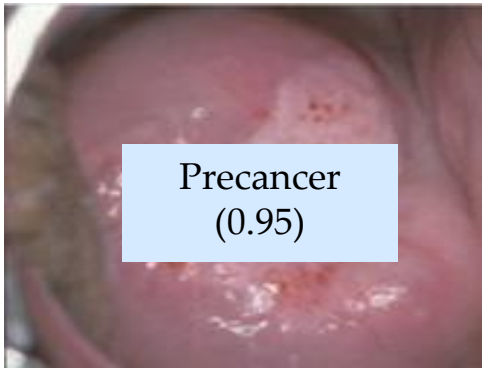
Zachary C. Lipton [1]

The false hope of current approaches to explainable artificial intelligence in health care

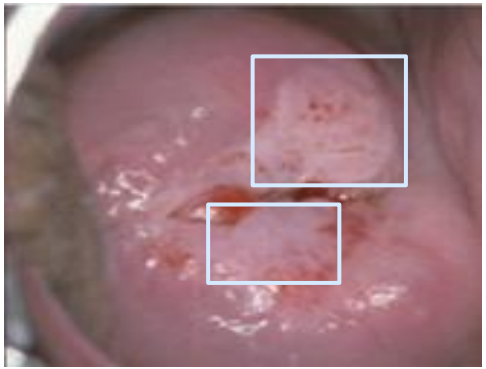
Marzyeh Ghassemi, Luke Oakden-Rayner, Andrew L Beam

Solution: Create models that are inherently more explainable

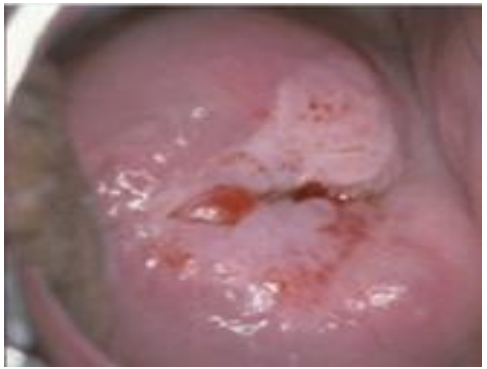
Classificat
ion task:



Detection
task:



Segmenta
tion task:



Classificat
ion tasks
can be
easiest to
annotate
for (can

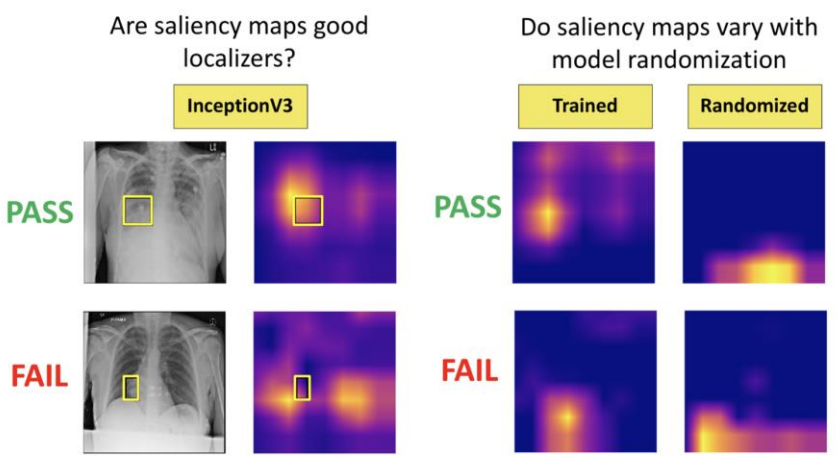
Problem 6: Evaluation plan

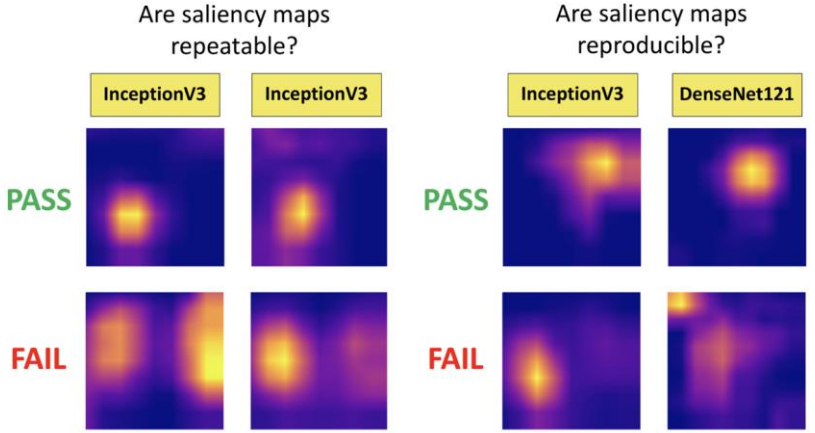
Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging

Nishanth Arun, BTech • Nathan Gaw, PhD* • Praveer Singh, PhD • Ken Chang, PhD • Mehak Aggarwal, MTech • Bryan Chen, MEng • Katharina Hoebel, MD • Sharut Gupta • Jay Patel, BS • Mishka Gidwani, BS • Julius Adebayo, MEng • Matthew D. Li, MD • Jayashree Kalpathy-Cramer, PhD*

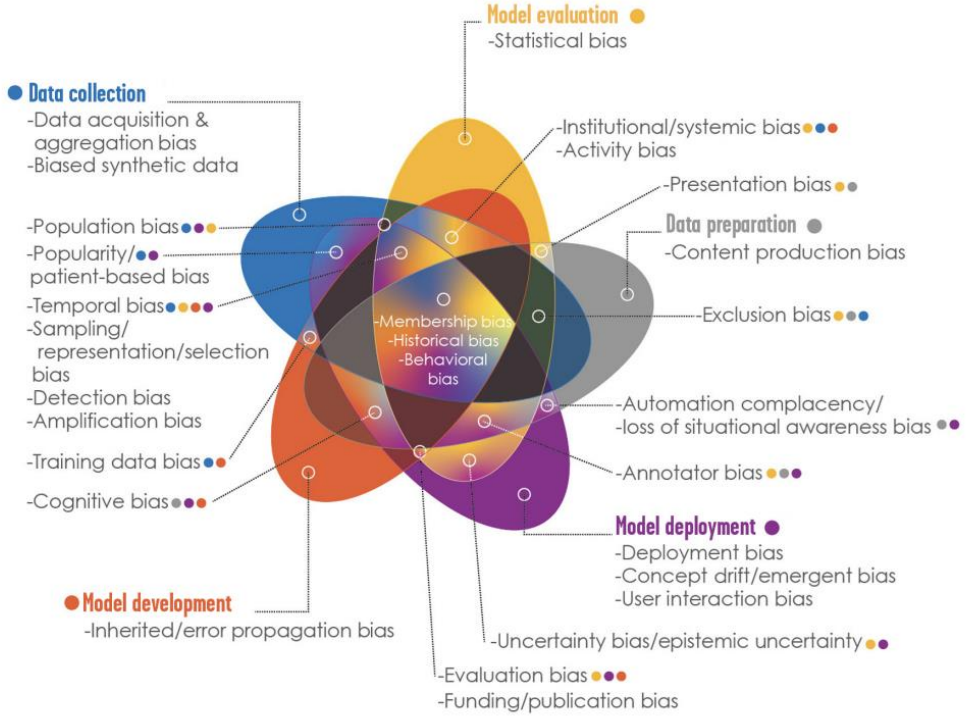




Algorithmic bias in medical image analysis

Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment

Karen Drukker[a,*] **Weijie Chen**[b] **Judy Gichoya**[c] **Nicholas Gruszauskas**[a] **Jayashree Kalpathy-Cramer**[d] **Sanmi Koyejo**[e] **Kyle Myers**[f] **Rui C. Sá**[g,h] **Berkman Sahiner**[b] **Heather Whitney**[a] **Zi Zhang**[i] **and Maryellen Giger**[a]



Problem 7: Machine learning models may be biased

NEWS · 24 OCTOBER 2019 · UPDATE 26 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals — and highlights ways to correct it.

ECONOMICS

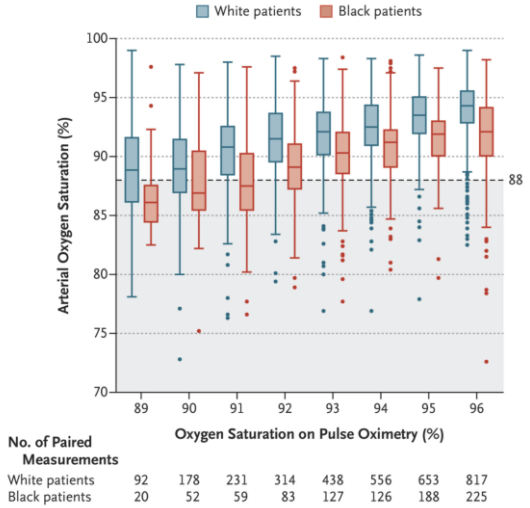
Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Problem 7: Potential Harm in the use of AI

AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind

Machine learning has the potential to save thousands of people from skin cancer each year—while putting others at greater risk.



Racial Bias in Pulse Oximetry Measurement

https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/

https://www.nejm.org/doi/10.1056/NEJMc2029240

Problem 7: **Models may be biased without us recognizing it!!!!**

- Learn to predict **self-reported racial** identity in medical images
- "models can be trained to predict race from medical images with high performance …x-ray imaging …AUC range 0·91–0·99"
- "Despite many attempts, we couldn't work out what it learns or how it does it. It didn't seem to rely on obvious confounders, nor did it rely on a limited anatomical region or portion of the image spectrum."

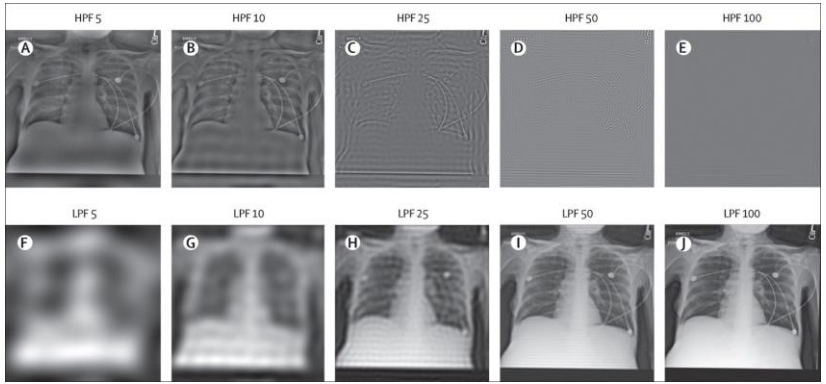
ARTICLES | VOLUME 4, ISSUE 6, E406-E414, JUNE 01, 2022

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, MD • Imon Banerjee, PhD • Ananth Reddy Bhimireddy, MS • John L Burns, MS • Leo Anthony Celi, MD • Li-Ching Chen, BS • et al. Show all authors

Open Access • Published: May 11, 2022 • DOI: https://doi.org/10.1016/S2589-7500(22)00063-2

Check for updates



https://laurenoakdenrayner.com/2021/08/02/ai-has-the-worst-superpower-medical-racism/

Problem 7: AI can be biased in ways that are hard to identify

This Issue | Views 1,589 | Citations 1 | Altmetric 212

Original Investigation

May 4, 2023

Association of Biomarker-Based Artificial Intelligence With Risk of Racial Bias in Retinal Images

Aaron S. Coyner, PhD[1]; Praveer Singh, PhD[2,3]; James M. Brown, PhD[4]; et al

Author Affiliations

JAMA Ophthalmol. 2023;141(6):543-552. doi:10.1001/jamaophthalmol.2023.1310

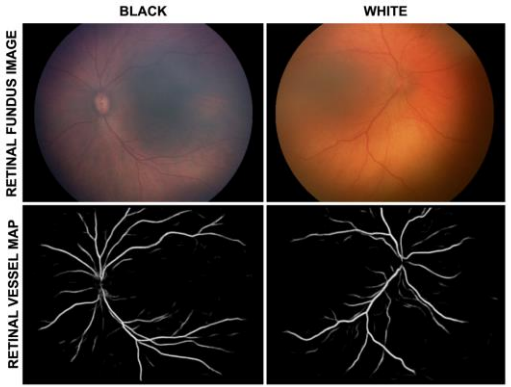
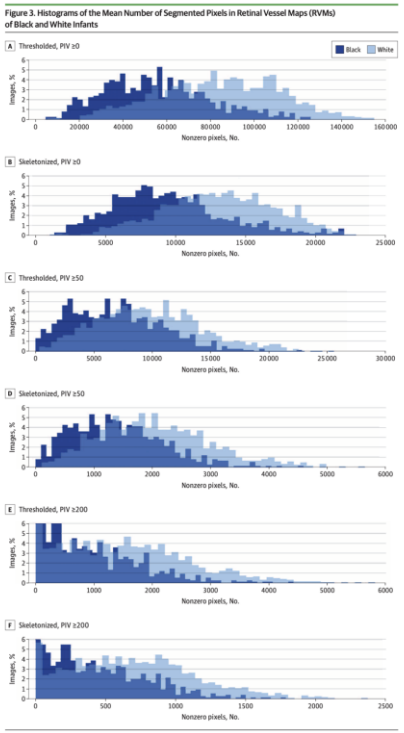


IMAGE TYPE	AUC-PR (image level)	AUC-ROC (image level)	AUC-PR (subject level)	AUC-ROC (subject level)
PIV ≥ 0				
Color RFI	0.999	0.999	1.000	1.000
Grayscale RVM	0.938	0.959	0.995	0.995
Binarized RVM	0.960	0.974	0.999	0.999
Skeletonized RVM	0.882	0.944	0.980	0.990



https://jamanetwork.com/journals/jamaophthalmology/article-abstract/2804442

Problem 7: Many datasets used to create AI lack diversity

September 22/29, 2020

Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms

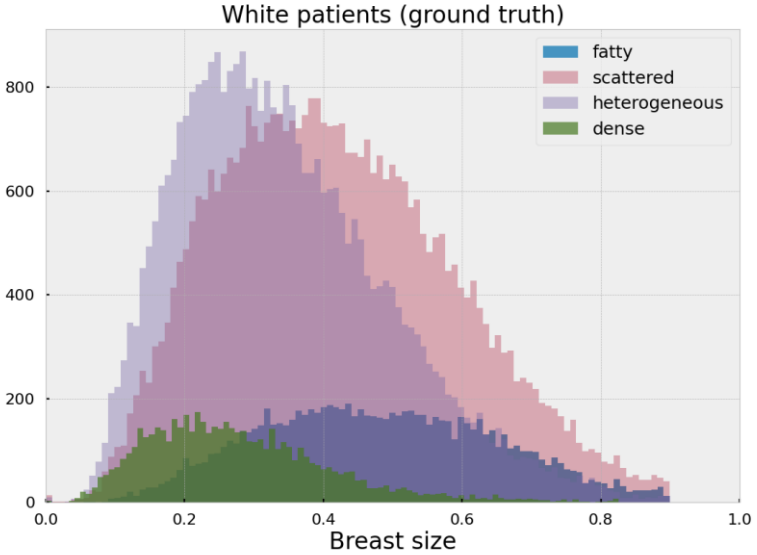
Amit Kaushal, MD, PhD[1]; Russ Altman, MD, PhD[1]; Curt Langlotz, MD, PhD[2]

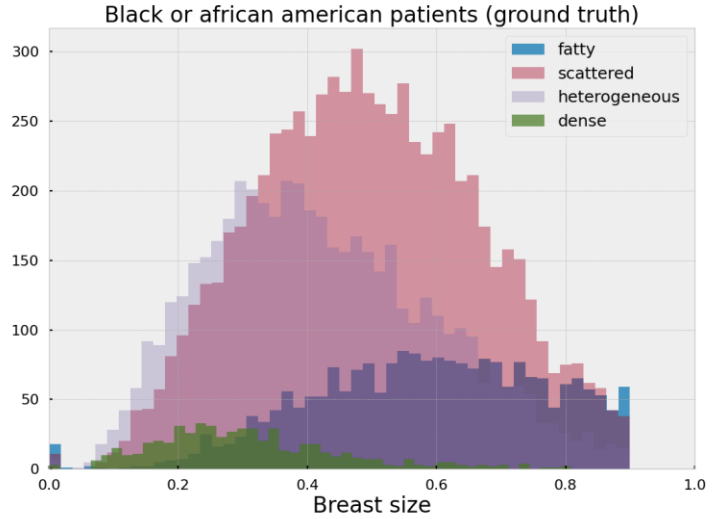
"In clinical applications of deep learning across multiple disciplines, algorithms trained on US patient data were disproportionately trained on cohorts from California, Massachusetts, and New York, with little to no representation from the remaining 47 states."

What is Fairness?

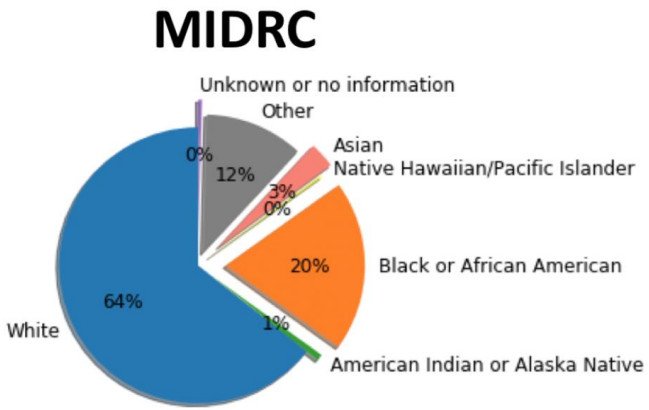
- *Fairness* is judged against set of ethical and legal principles, which can change over time and vary between groups, cultures, countries
- Fairness usually considered on an individual or group level
  - Individual fairness - similar individuals treated similarly
  - Group fairness - different groups treated equally
- To quantify unfairness, several mathematical definitions of fairness exist

Achieving Fairness can be challenging



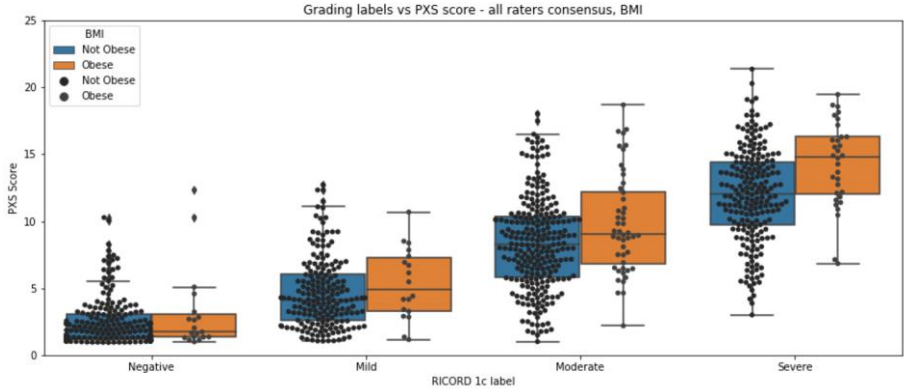


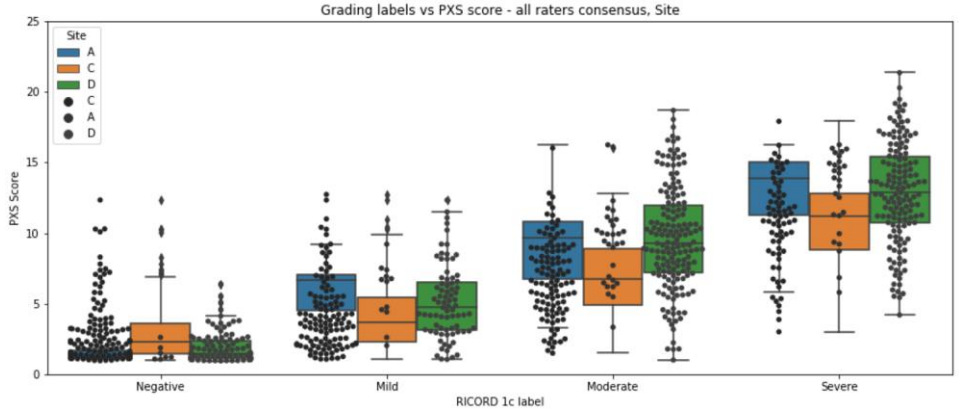
Solution 7: Increase diversity of datasets, measure in all populations



Diversity in the MIDRC dataset (May 2021).

https://www.midrc.org/diversity





Problem 7: Evaluation plan

Evaluation model performance in sub-populations

Evaluate failure cases to better understand sub-populations to study

Study if “shortcut learning” is occuring

Does AI have super-human capabilities?

Article | Published: 19 February 2018

Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning

Ryan Poplin, Avinash V. Varadarajan, Katy Blumer, Yun Liu, Michael V. McConnell, Greg S. Corrado, Lily Peng & Dale R. Webster

Nature Biomedical Engineering **2**, 158–164 (2018) | Cite this article

22k Accesses | **653** Citations | **2388** Altmetric | Metrics

Predicting sex from retinal fundus photographs using automated deep learning

Edward Korot[1], Nikolas Pontikos[1], Xiaoxuan Liu[1,2,3], Siegfried K. Wagner[1], Livia Faes[1,4], Josef Huemer[1,5], Konstantinos Balaskas[1], Alastair K. Denniston[1,2,3,6], Anthony Khawaja[1] & Pearse A. Keane[1]

Predicting risk of breast cancer at one to five years from the mammogram.

Retinal microvasculature dysfunction is associated with Alzheimer's disease and mild cognitive impairment

Check for updates

Jacqueline Chua[1,2,3], Qinglan Hu[1,3], Mengyuan Ke[1,3], Bingyao Tan[1,3,4], Jimmy Hong[1], Xinwen Yao[1,3,4], Saima Hilal[5,6,7], Narayanaswamy Venketasubramanian[5,8], Gerhard Garhöfer[9], Carol Y. Cheung[10], Tien Yin Wong[1,2], Christopher Li-Hsian Chen[5] and Leopold Schmetterer[1,2,3,4,9,11,12*]

ORIGINAL REPORTS | Breast Cancer

Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model

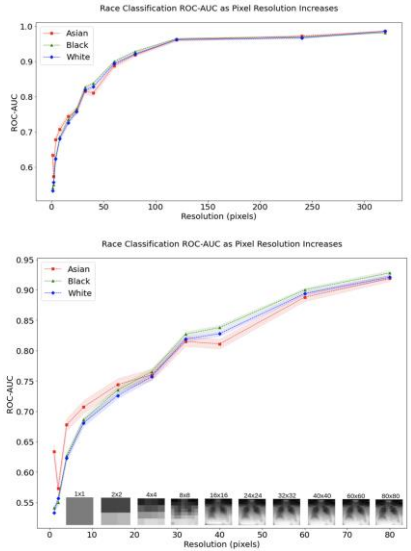
Check for updates

Adam Yala, MEng[1,2]; Peter G. Mikhael, BS[1,2]; Fredrik Strand, MD, PhD[3,4]; Gigin Lin, MD, PhD[5]; Siddharth Satuluru, BS[6]; Thomas Kim, MS[7]; ...

Superhuman + risk of bias + not transparent –> need for vigilance?

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, Haoran Zhang

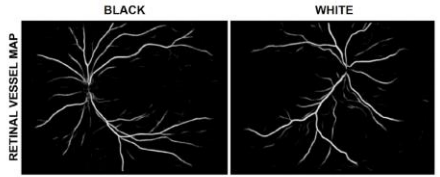


Not Color Blind: AI Predicts Racial Identity from Black and White Retinal Vessel Segmentations

Aaron S Coyner PhD[1,a], Praveer Singh PhD[2,3,a], James M Brown, PhD[4], Susan Ostmo MS[1], RV Paul Chan MD[5], Michael F Chiang MD, MA[6], Jayashree Kalpathy-Cramer PhD[2,3,b], J Peter Campbell MD, MPH[1,b]

Surprisingly:

Grayscale Retinal Vessel Maps Contain Information Associated with Self-Reported Race



	AUC-PR (image level)	AUC-ROC (image level)	AUC-PR (subject level)	AUC-ROC (subject level)
Grayscale RVM	0.938	0.959	0.995	0.995

Grayscale Retinal Vessel Maps Are Associated with Self-Reported Race

Implications for Artificial Intelligence Models

Aaron S. Coyner, PhD

OHSU

Just hear me out…

Tom Yankeelov

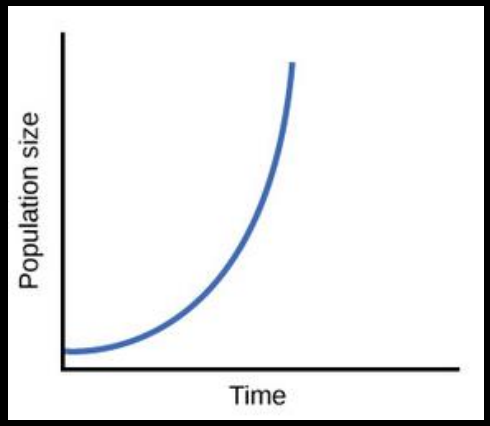
The University of Texas at Austin

2 October 2023

Building a mechanism-based model

Exponential Growth

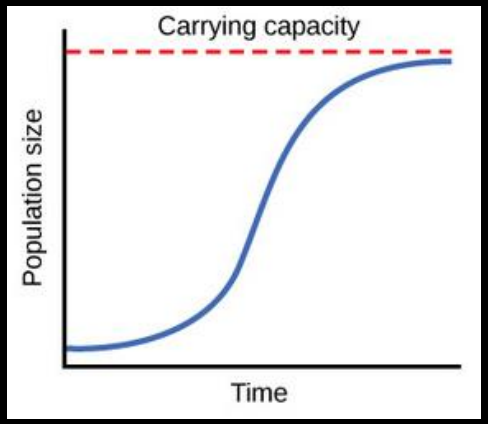
$$\frac{dN}{dt} = kN \quad \Rightarrow \quad N(t) = N_0 \exp(k \times t)$$



Parameters

$N(t)$ = # of tumor cells at time t
N_0 = initial # of tumor cells
k = proliferation rate

Building a mechanism-based model



Exponential Growth

$$\frac{dN}{dt} = kN \quad \Rightarrow \quad N(t) = N_0 \exp(k \cdot t)$$

Logistic Growth

$$\frac{dN}{dt} = kN\left(1 - \frac{N}{\theta}\right) \quad \Rightarrow N(t) = \frac{\theta N_0}{N_0 + (\theta - N_0)\exp(-k \cdot t)}$$

Parameters

$N(t)$ = # of tumor cells at time t
N_0 = initial # of tumor cells
k = proliferation rate
θ = carrying capacity

→ Now need to account for spatial variations in tumor growth

Reaction-diffusion equation

$$\frac{\partial N}{\partial t} = \nabla \bullet (D \nabla N) + kN\left(1 - \frac{N}{\theta}\right)$$

$N(t)$ = # of tumor cells at time t
k = proliferation rate
θ = carrying capacity
D = tumor cell diffusion

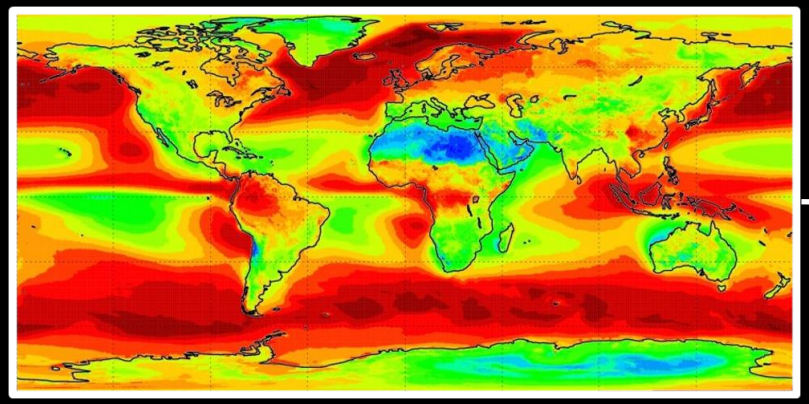
Building a mechanism-based model

→ And if we include the effects of therapy:

$$\frac{\partial N(x,t)}{\partial t} = \underbrace{\nabla \cdot (D \nabla N(x,t))}_{\textbf{Movement}} + \underbrace{k(x) N(x,t) \left(1 - \frac{N(x,t)}{\theta}\right)}_{\textbf{Proliferation}} - \underbrace{N(x,t) \sum_{n} \alpha_n \exp(\beta_n t) C_{tissue}^{drug,n}(x,t)}_{\textbf{Treatment}}$$

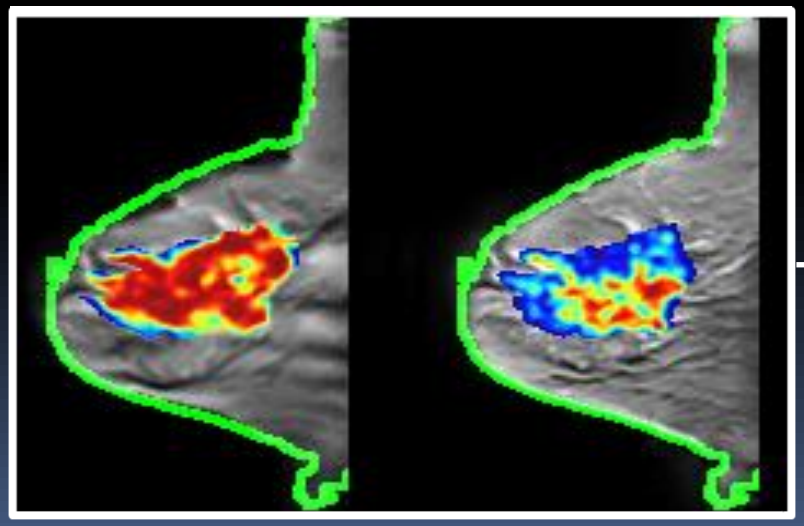
So, what do we do with this model?

Jarrett et al., Nature Protocols, 2021



Weather model → Weather forecast

Just as satellites provide the data for weather forecasting, quantitative imaging data can provide the data for tumor forecasting.



Tumor model → Tumor forecast

Yankeelov, Quaranta, Evans, Rericha. Cancer Research, 2015

Applying a mechanism-based model

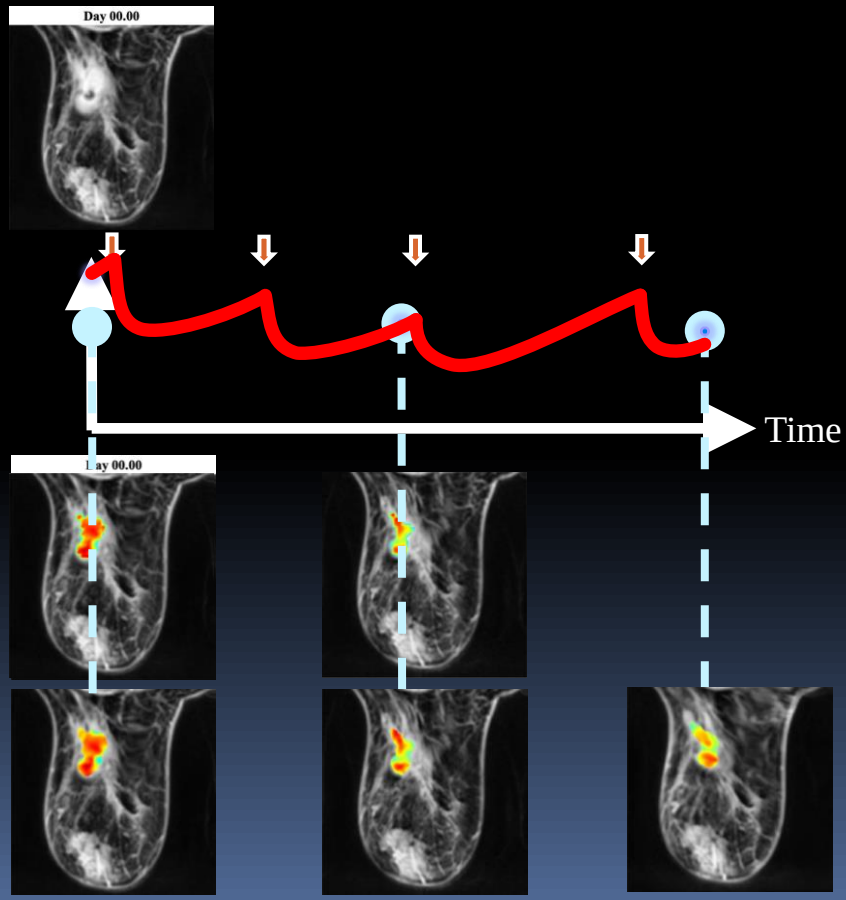
$$\frac{\partial N(x,t)}{\partial t} = \underbrace{\nabla \cdot (D \nabla N(x,t))}_{\text{Movement}} + \underbrace{k(x) N(x,t) \left(1 - \frac{N(x,t)}{\theta}\right)}_{\text{Proliferation}} - \underbrace{N(x,t) \sum_n \alpha_n \exp(\beta_n t) C_{tissue}^{drug,n}(x,t)}_{\text{Treatment}}$$

Spatially-resolved drug kinetics

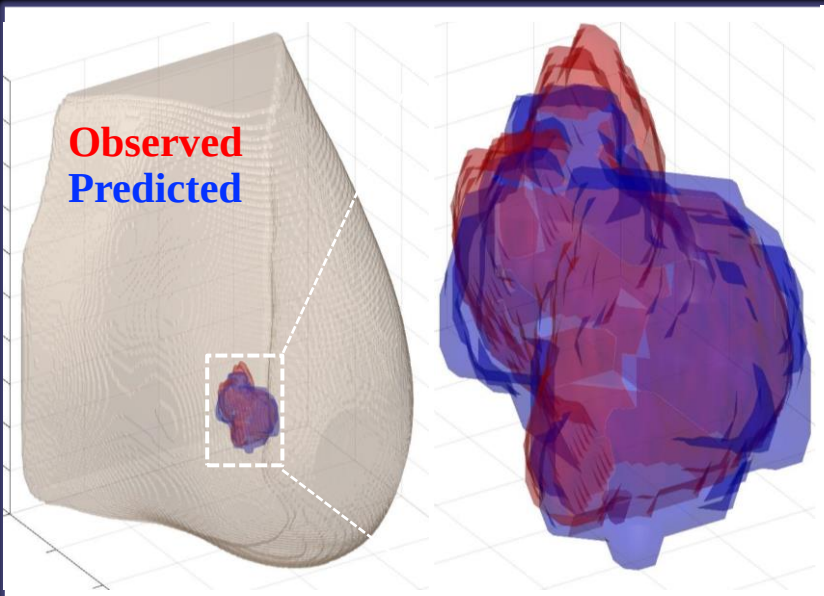
Overall cellularity

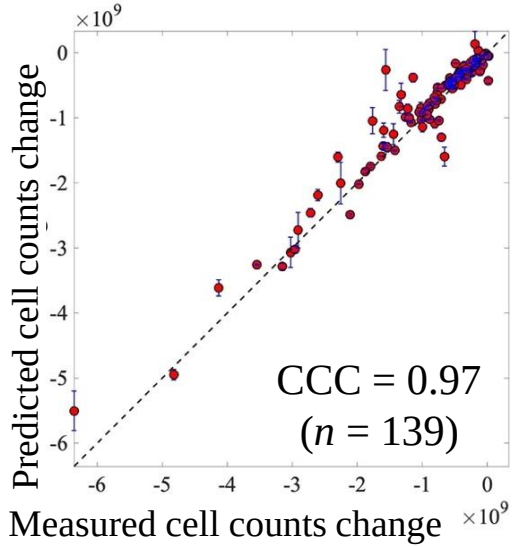
Predict spatio-temporal cellularity

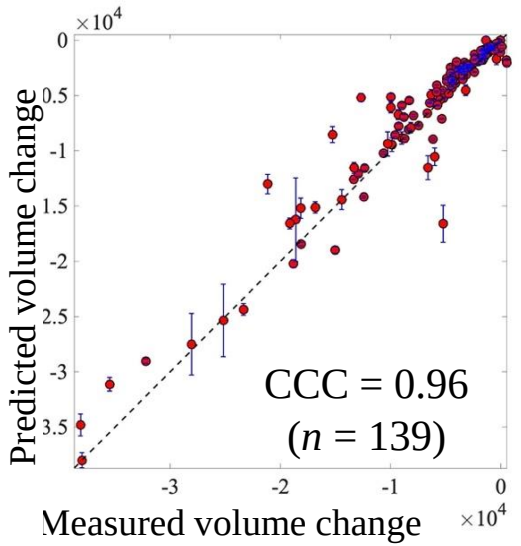
Measure spatio-temporal cellularity



Mechanism-based models enable patient specific predictions





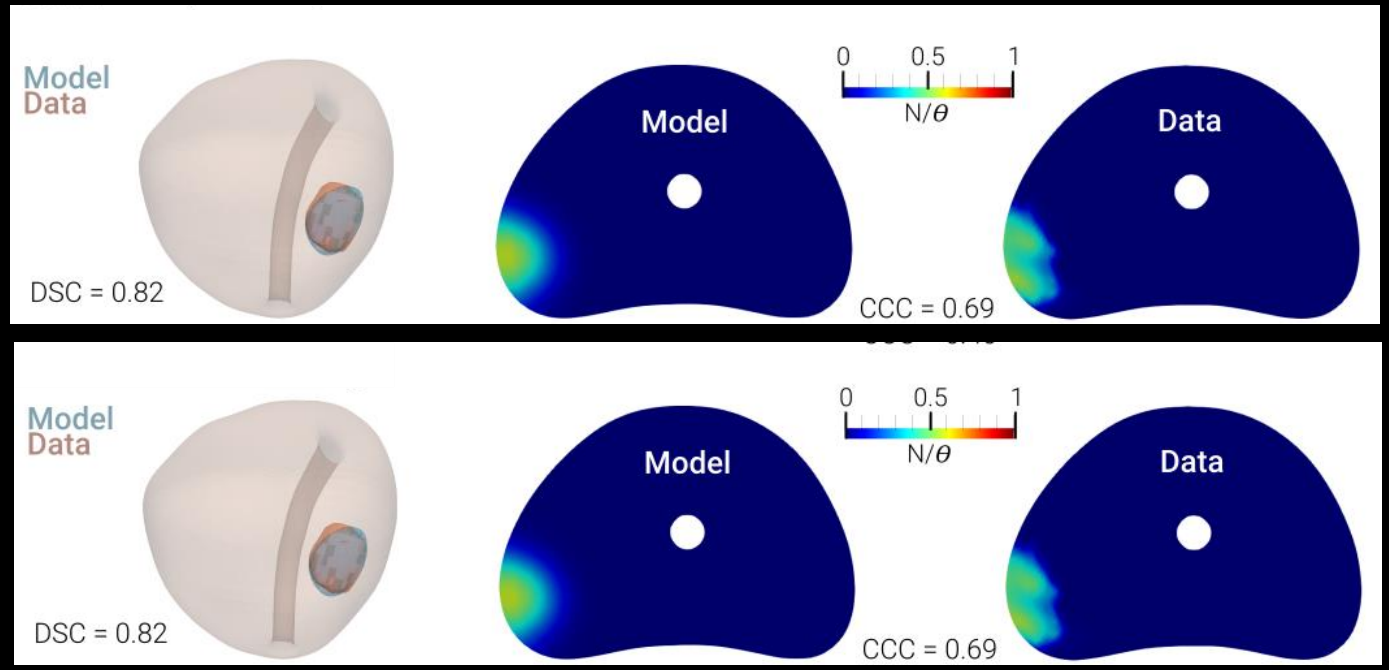


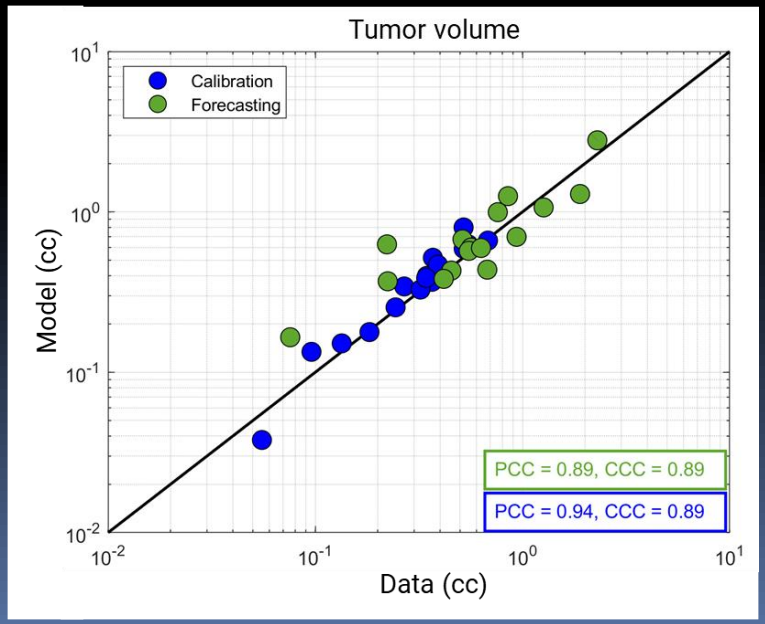
AUC = 0.89
(*n* = 50)

→ *We are getting pretty good at predicting the spatial and temporal development of these breast tumors in the neoadjuvant setting…*

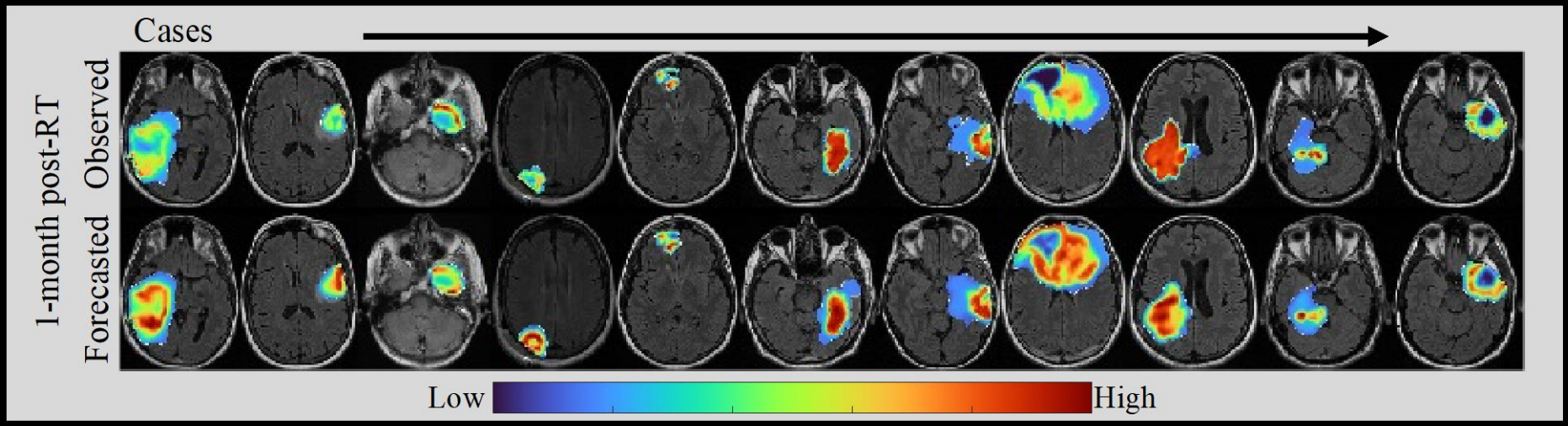
Wu, et al. Cancer Research, 2022

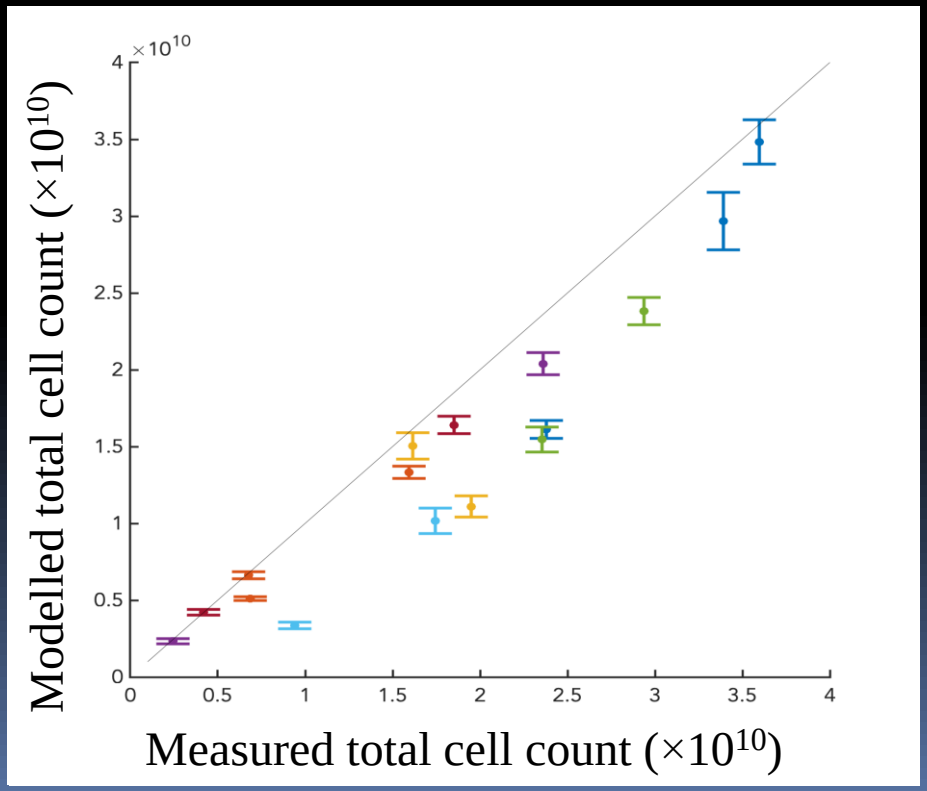
… with similar results for prostate cancer…





… and for brain cancer





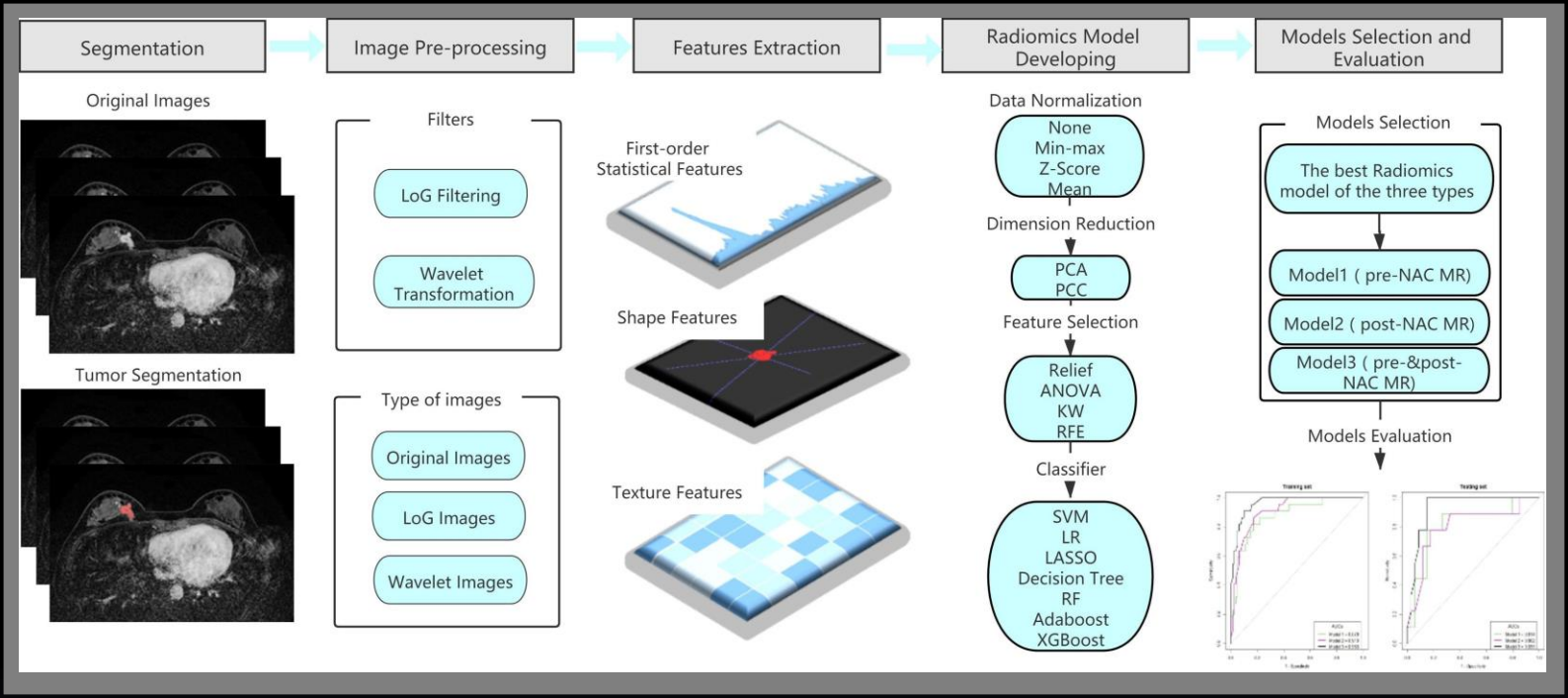
I know what you are thinking

"Why did you drag me through all that math…

… isn't AI/Big Data just going to figure it all out?"

AI & Big Data… because who needs science?

- **Study goal:** Establish radiomics prediction models based on MRI for predicting recurrence of TNBC patients (*n* = 147) after NAT



→ 102 radiomics features were extracted and three models built based on:

1) pre-NAT MRI features ….. 0.81
2) post-NAT MRI features ….. 0.80
3) pre- and post-NAT MRI features ….. 0.93

} Area under ROC curve

Ma et al. Eur J Radiol. 2022;146:110095

Problem solved, right?

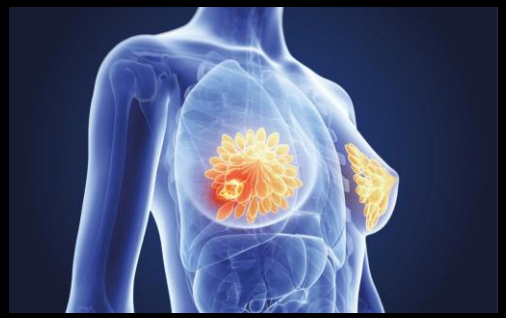
Well…

Let’s contrast this with a mechanism-based approach

Digital twin for predicting/optimizing treatment response

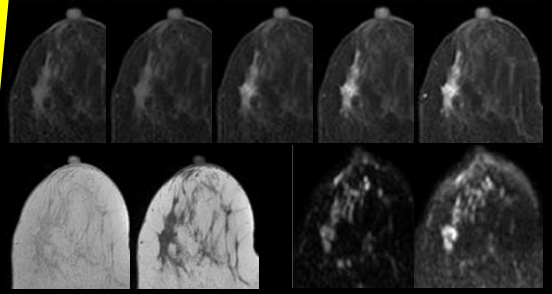
Physical state, S_i

Anatomy & morphology,
mechanical & physiological state



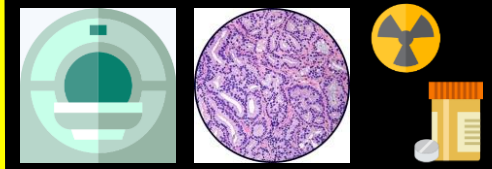
Observational data, O_i

Anatomy, perfusion, permeability,
cell density, metabolism



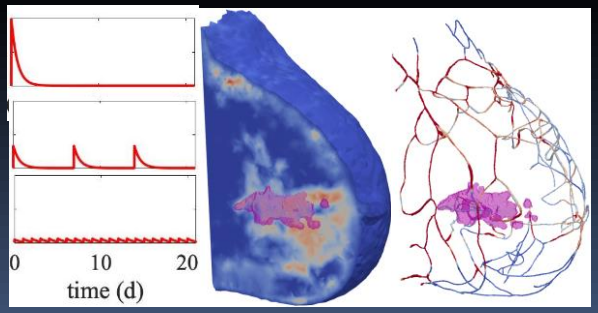
Control inputs, U_i

MRI studies
Biopsies
Optimize treatment



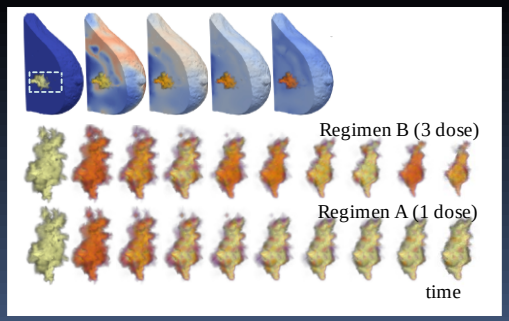
Digital state, D_i

Domain: FE mesh, boundary conditions
Parameters: tumor dynamics, mechanics
Inputs: treatment regimens



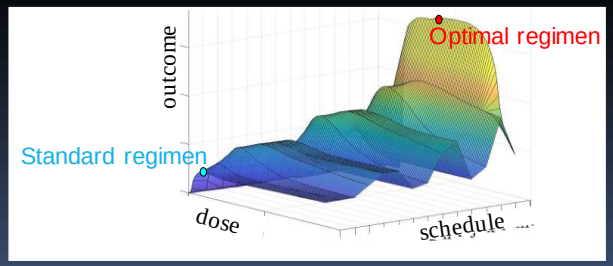
Quantities of interest, Q_i

Distribution of therapies,
tumor shape, cell density



Rewards, R_i

Outcomes; e.g.,
treatment efficacy & toxicity



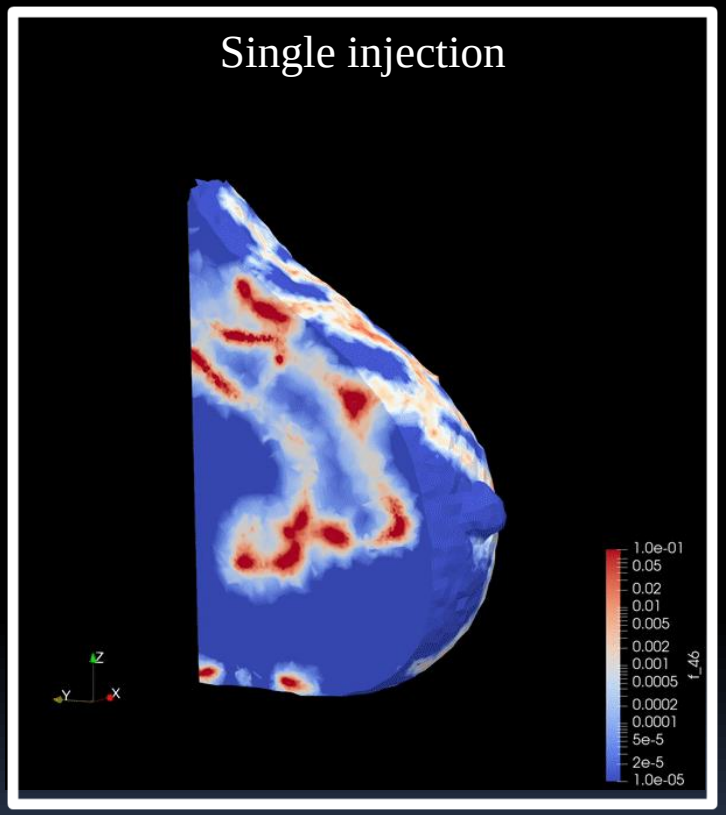
Physical asset

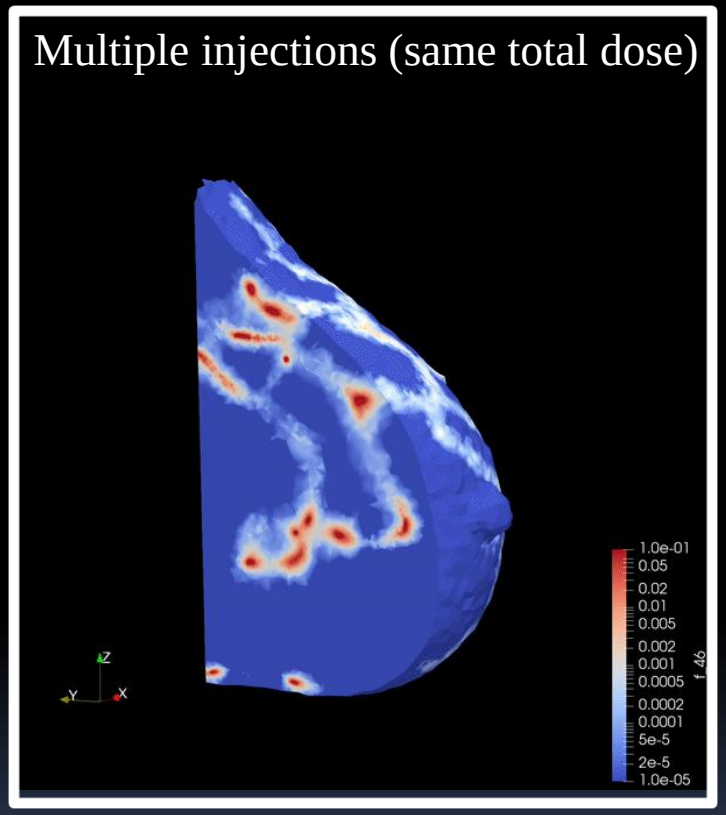
Digital asset

Wu, Lorenzo, Hormuth, Lima, et al. Biophysics Rev, 2022.

Digital twin for predicting/optimizing treatment response

Want to not just make predictions; want to optimize outcomes

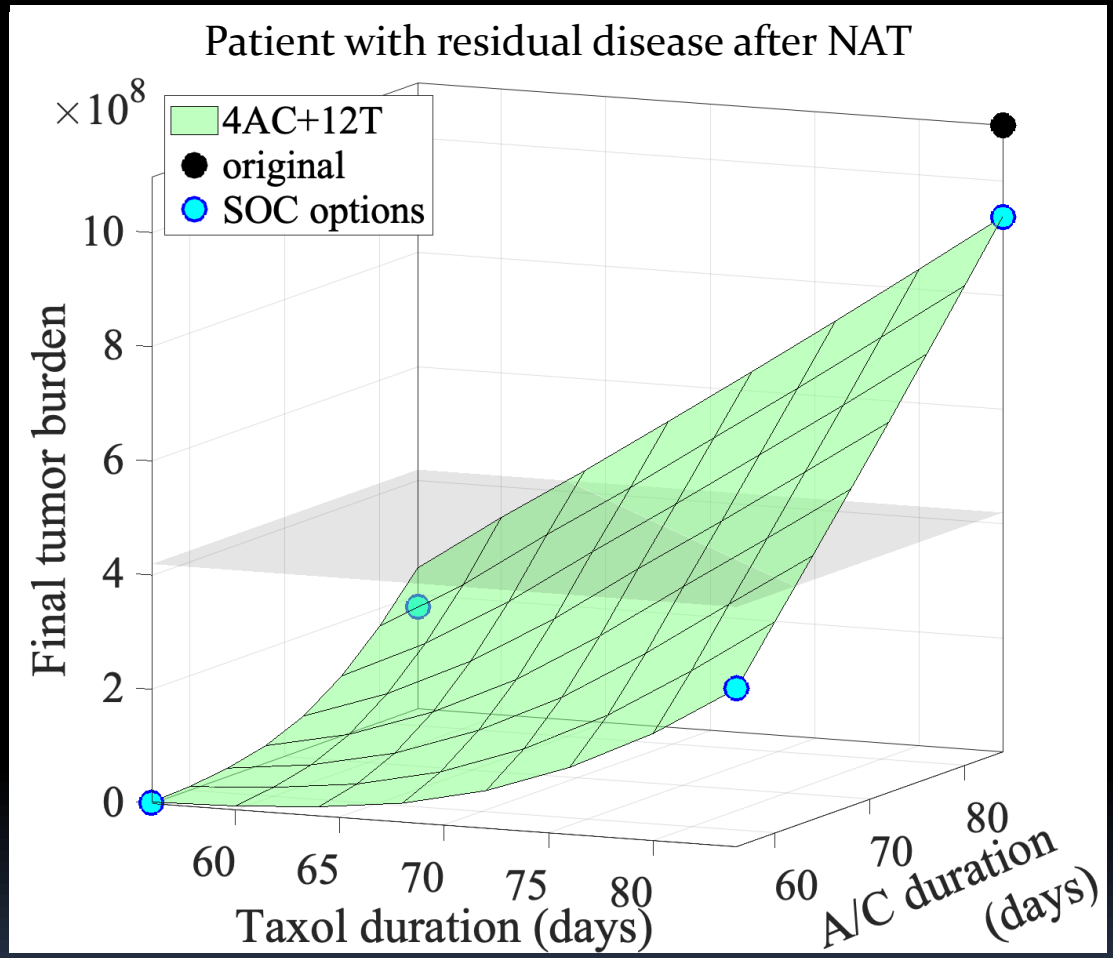




→ This formalism allows you to identify treatment protocols that balance treatment efficacy and safety

Chengyue Wu, et al. IEEE-TMI, 2020; IEEE TBME, 2022.

Digital twin for predicting/optimizing treatment response



You cannot do this with AI/Big Data only approach

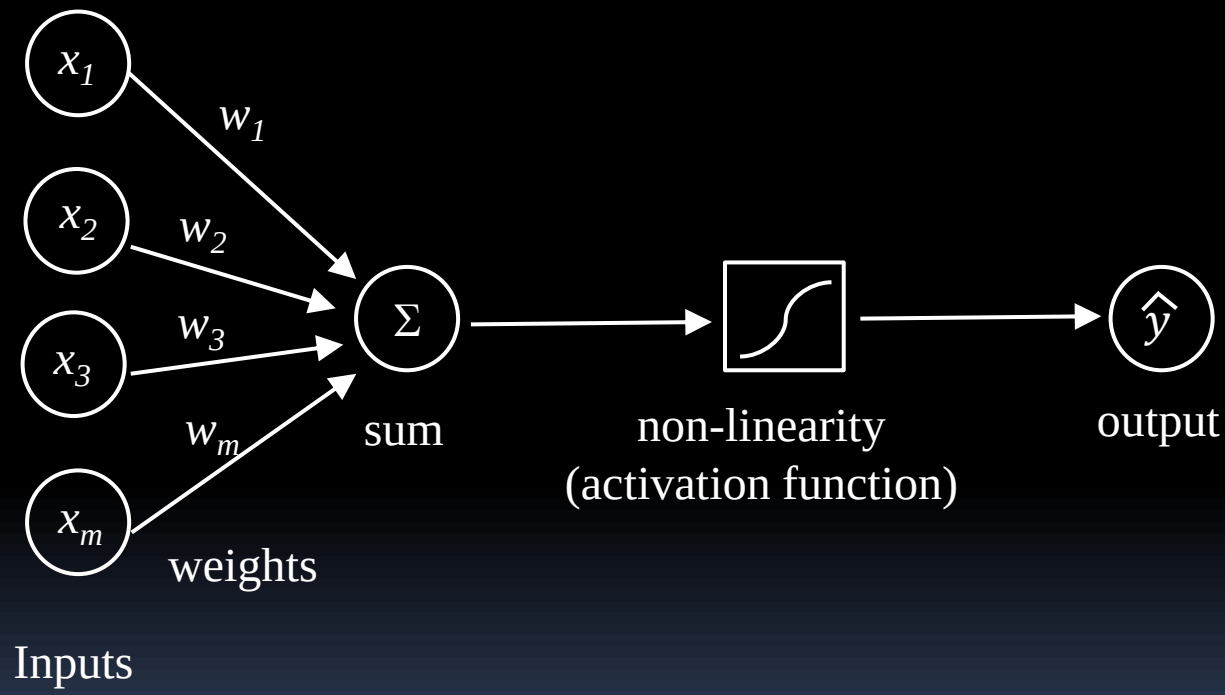
Must have a mechanism-based model

Chengyue Wu, et al. SABCS 2023.

Let’s take a deeper dive into deep learning to see why…

Quick peek into the guts of DL

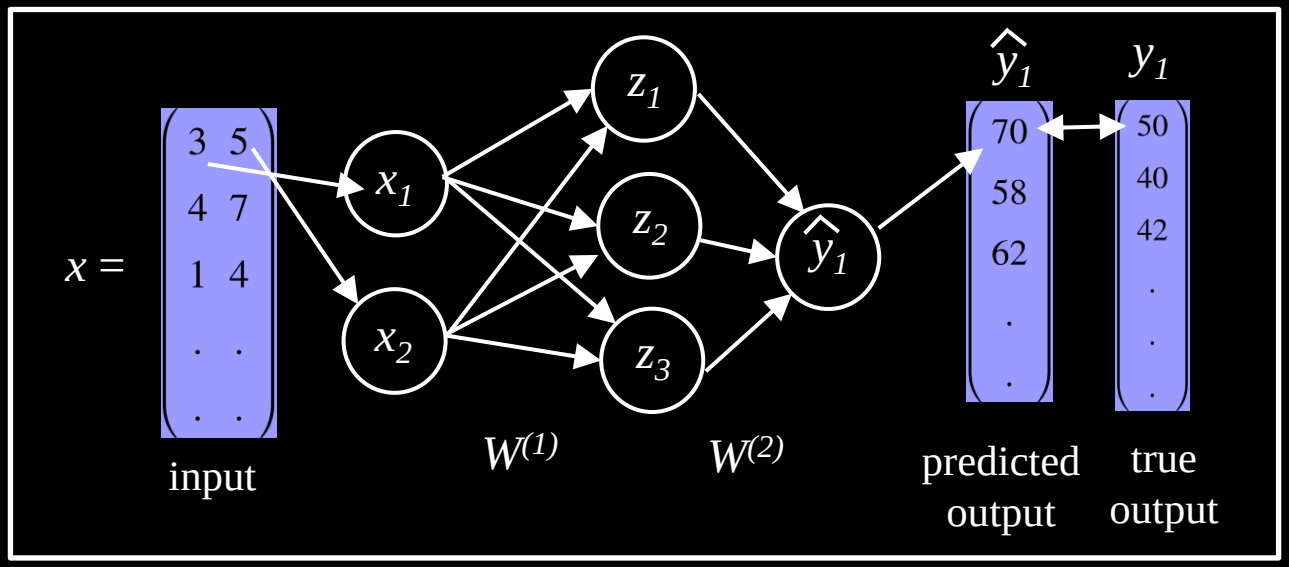
- Building block of DL is the "perceptron"; it takes some input data and maps it to output:



Alexander Amini; https://www.youtube.com/watch?v=7sB052Pz0sQ&list=PLtBw6njQRU-rwp5__7C0oIVt26ZgjG9NI

Quick peek into the guts of DL

- So this is what you do with this thing:



→ And you try to minimize something like the following by getting the best set of weights:

$$J(W) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i, w)\right)^2$$

Predicted = $\hat{y}_1$

→ We need LOTS of data to "train" the DL model; i.e., to calibrate the *w*'s

→ The "deeper" the neural network, the more data you need to train the network

Alexander Amini; https://www.youtube.com/watch?v=7sB052Pz0sQ&list=PLtBw6njQRU-rwp5__7C0oIVt26ZgjG9NI

But that training set does not exist for a host of problems…

In fact, we have already thrown AI/Big Data at cancer…


$0
$1,200
$0
KEN
WATSON
ДУМАЙ
Brad

A cautionary tale

→ From the IBM website:

"Watson for Oncology combines leading oncologists' deep expertise in cancer care with the speed of IBM Watson to help clinicians as they consider individualized cancer treatments for their patients."

A cautionary tale

→ From the IBM website:

"Watson for Oncology combines leading oncologists' deep expertise in cancer care with the speed of IBM Watson to help clinicians as they consider individualized cancer treatments for their patients."

→ From STAT in 2017:

"IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show"

A cautionary tale

→ From the IBM website:

"Watson for Oncology combines leading oncologists' deep expertise in cancer care with the speed of IBM Watson to help clinicians as they consider individualized cancer treatments for their patients."

→ From STAT in 2017:

"IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show"

→ From FORBES in 2017/2018:

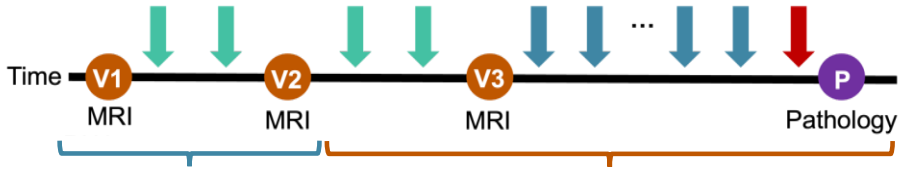
"IBM announced that its Watson Health chief had stepped down" and "its once-hyped A.I. business has been scaled back with layoffs"

Is there another way forward?

Linking mechanism-based and data-based modeling

Methods

Mechanism-based modeling



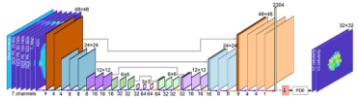
Calibration

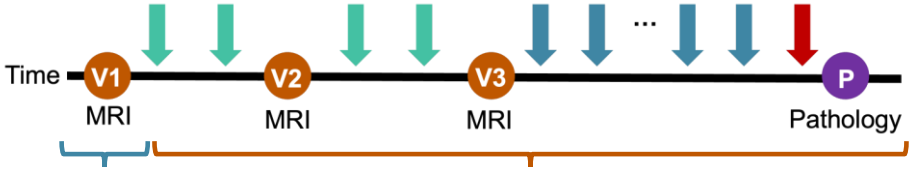
Predict response

Model parameters

Deep learning

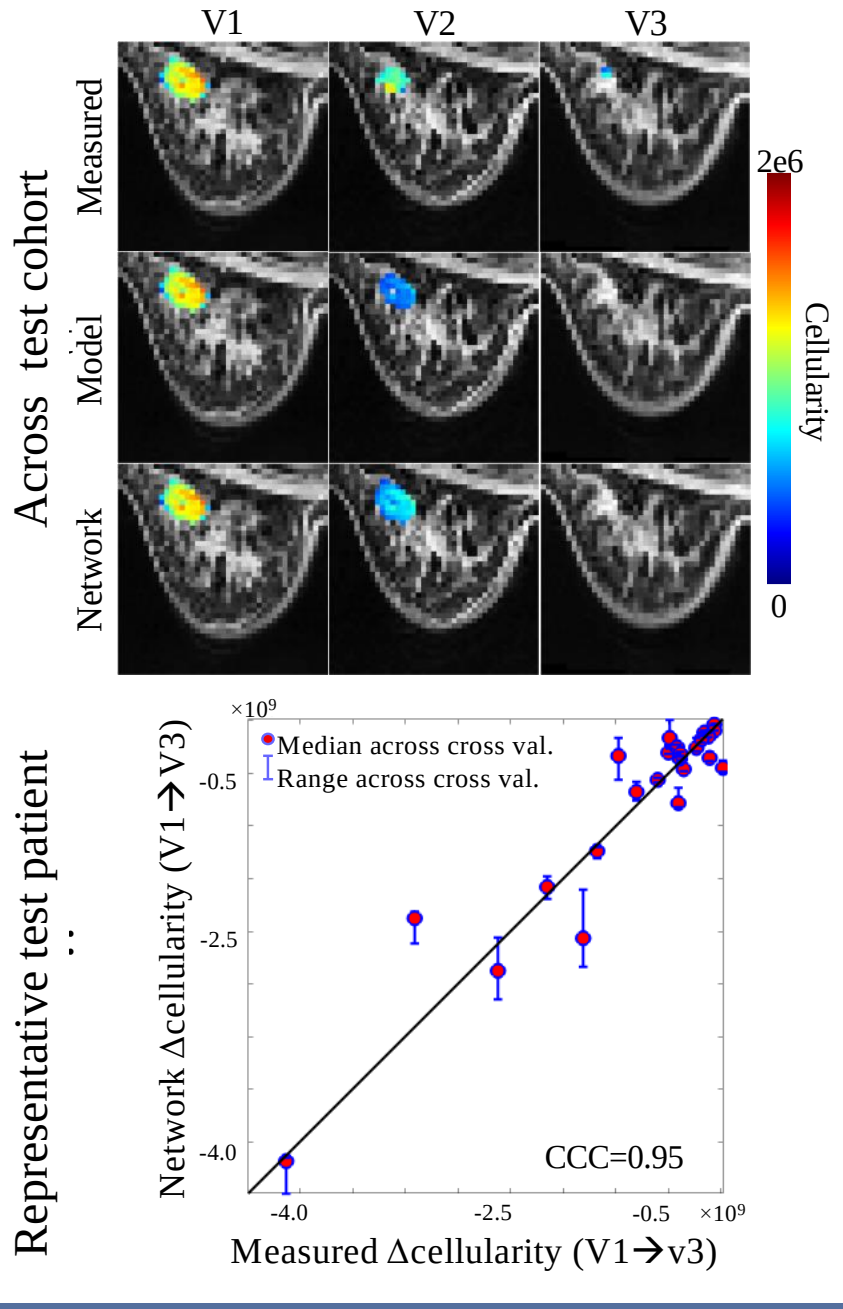
Train network





Input

Prediction: parameters & response



Casey Stowers, et al. SABCS 2023.

A Plea

- Statistical inference—though, enormously powerful—relies on properties of large populations that obscure conditions specific to the individual

- High-consequence decisions (e.g., those in oncology) must be based on more than just data analytics

→ These decisions must incorporate biophysical processes that can be calibrated with patient-specific data to make patient-specific predictions

If you want to design something that is useful for an individual human being, you must rely on that human being's unique characteristics

- So, build your neural networks if you must…

… but please don't forget about $F = ma$

Thank you very much
for your time and attention.

@UTCompOnco
cco.oden.utexas.edu
tey@utexas.edu